nature structural & molecular biology

Weak seed-pairing stability and high target-site abundance decrease the proficiency of *lsy-6* and other microRNAs

David M Garcia[1–3,8], Daehyun Baek[1–5,8], Chanseok Shin[1–3,6], George W Bell[1], Andrew Grimson[1–3,7] & David P Bartel[1–3]

Most metazoan microRNAs (miRNAs) target many genes for repression, but the nematode *lsy-6* miRNA is much less proficient. Here we show that the low proficiency of *lsy-6* can be recapitulated in HeLa cells and that miR-23, a mammalian miRNA, also has low proficiency in these cells. Reporter results and array data indicate two properties of these miRNAs that impart low proficiency: their weak predicted seed-pairing stability (SPS) and their high target-site abundance (TA). These two properties also explain differential propensities of small interfering RNAs (siRNAs) to repress unintended targets. Using these insights, we expand the TargetScan tool for quantitatively predicting miRNA regulation (and siRNA off-targeting) to model differential miRNA (and siRNA) proficiencies, thereby improving prediction performance. We propose that siRNAs designed to have both weaker SPS and higher TA will have fewer off-targets without compromised on-target activity.

MicroRNAs are ~22-nucleotide (nt) RNAs that pair with the messages of protein-coding genes to direct post-transcriptional repression of these target mRNAs[1,2]. In animals, many studies using a wide range of methods, including comparative sequence analysis, site-directed mutagenesis, genetics, mRNA profiling, coimmunoprecipitation and proteomics, have shown that perfect pairing with miRNA nucleotides 2–7, known as the miRNA seed, is important for the recognition of many miRNA targets[3]. To impart more than marginal repression of mammalian targets, this seed pairing is usually augmented by either a match with miRNA nucleotide 8 (7-mer-m8 site)[4–7], an A across from nucleotide 1 (7-mer-A1 site)[4,7] or both (8-mer site)[4,7]. In rare instances, targeting also occurs through 3′-compensatory sites[4,5,8] and centered sites[9], for which substantial pairing outside the seed region compensates for imperfect seed pairing.

A single miRNA can target hundreds of distinct mRNAs through seed-matched sites[10]. Indeed, most human mRNAs are conserved regulatory targets[8], and many additional regulatory interactions occur through nonconserved sites[11–13]. However, not every site is effective; 8-nt sites are effective more often than 7-nt sites, which are effective more often than 6-nt sites[7,14]. Another factor is site context. For example, sites in the 3′ untranslated regions (3′ UTRs) are effective more often than those in the path of the ribosome[7]. Among 3′ UTR sites, those away from the centers of long UTRs and those within high local A-U sequence context are effective more often[7], consistent with reports that sites predicted to be within more accessible secondary structure tend to be more effective[15–19]. Site efficacy is also influenced by proximity to other miRNA-binding sites[7,20], to protein-binding sites[21] and to sequences that can pair with the 3′ region of the miRNA, particularly nucleotides 13–17 (ref. 7).

Studies of site efficacy have focused primarily on different sites for the same miRNA, without systematic investigation of whether some miRNA sequences are more proficient at targeting than others. Broadly conserved miRNAs typically have many more conserved targeting interactions than do other miRNAs[4,8], and highly or broadly expressed miRNAs seem to target more mRNAs than do others[22], but these phenomena reflect evolutionary happenstance more than intrinsic targeting proficiency.

Our interest in targeting proficiency was spurred by results regarding the *lsy-6* miRNA. When tested in *Caenorhabditis elegans*, only 1 of 14 predicted targets with 7- to 8-nt seed matched sites responds to *lsy-6*, which was interpreted to show that perfect seed pairing is not a reliable predictor for miRNA-target interactions[23]. Alternatively, and in keeping with findings for many other miRNAs[3], the results for *lsy-6* might not apply to other miRNAs because *lsy-6* could have unusually high targeting specificity owing to unusually low targeting proficiency. A similar rationale might explain results for mammalian miR-23, another miRNA that confers unusually weak responses from most reporters designed to test predicted targets.

When considering properties that might confer a low targeting proficiency, we noted that both *lsy-6* and miR-23 have unusually (A+U)-rich seed regions, which could lower the stability of seed-pairing interactions. Perhaps a threshold of SPS is required for the miRNA to remain associated with targets long enough to achieve widespread seed-based targeting. Indeed, predicted SPS is correlated with the propensity of siRNAs to repress unintended targets[24], a process

[1]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA. [2]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [4]School of Biological Sciences, Seoul National University, Seoul, Republic of Korea. [5]Bioinformatics Institute, Seoul National University, Seoul, Republic of Korea. [6]Department of Agricultural Biotechnology, Seoul National University, Seoul, Republic of Korea. [7]Present address: Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York, USA. [8]These authors contributed equally to this work. Correspondence should be addressed to D.P.B. (dbartel@wi.mit.edu) and D.B. (baek@snu.ac.kr).

Received 12 December 2010; accepted 1 July 2011; published online 11 September 2011; doi:10.1038/nsmb.2115

called "off-targeting," which occurs through the same seed-based recognition as that for endogenous miRNA targeting[10]. Potentially confounding this interpretation, however, miRNAs with (A+U)-rich seed regions have more 3′ UTR–binding sites, a consequence of the (A+U)-rich nucleotide composition of 3′ UTRs, which could dilute the effect on each target message. Indeed, TA can be manipulated to titrate miRNAs away from their normal targets[25,26], and natural TA has been proposed to influence miRNA targeting and siRNA off-targeting[27,28], although these reported TA effects have not been fully disentangled from potential SPS effects. Here, we find that both SPS and TA have a substantial impact on targeting proficiency, and apply these insights to improve miRNA target predictions.

RESULTS

lsy-6 targeting specificity is recapitulated in HeLa cells

lsy-6 targeting was originally examined in a *C. elegans* neuron[23], whereas more proficient targeting by other miRNAs has been experimentally demonstrated in other systems, sometimes in vertebrate tissues or primary cells[11,13,29,30] but more often in cell lines[3]. To test whether differences in targeting proficiency could be attributed to the different biological contexts in which the miRNAs had been examined, we ported the 14 3′ UTRs tested in *C. elegans* into a luciferase reporter system typically used in mammalian cell lines and introduced the *lsy-6* miRNA by co-transfecting an imperfect RNA duplex representing the miRNA (**Fig. 1a**) and the short RNA from the other arm of the hairpin, known as the miRNA* (**Supplementary Fig. 1a**). As has been observed in worms[23], only the *cog-1* 3′ UTR responded in HeLa cells (**Fig. 1b**). Repression was lost when a control miRNA (miR-1) replaced *lsy-6* or when the two *cog-1* sites were mutated, introducing either mismatches (**Fig. 1b**) or G•U wobbles (**Supplementary Fig. 1b,c**).

Each of the 14 3′ UTRs had at least one canonical 7- to 8-nt *lsy-6* site, and 11 UTRs had a site conserved in three sequenced nematodes (**Supplementary Table 1**). When evaluated using the context-score model, some sites had scores comparable to those of sites that mediate repression in this assay[7] (**Supplementary Table 1**). Moreover, the C27H6.9 3′ UTR had two 8-mer sites with scores matching those of the two *cog-1* sites. The close match between the results in our heterologous reporter assay and previous results in *C. elegans* neurons indicated that the specificity for targeting the *cog-1* 3′ UTR did not require the endogenous cellular context of *lsy-6* repression; it was operable in HeLa cell culture and thereby attributable to the intrinsic properties of *lsy-6* and its targets. This result also indicated that these properties could be investigated in mammalian cell culture, which is easier than using stable reporter lines in worms.

Modifying both SPS and TA elevates targeting proficiency

As expected for a miRNA with sequence UUUGUAU at nucleotides 2–8, the calculated free energy (ΔG°) of the predicted SPS for the *lsy-6*

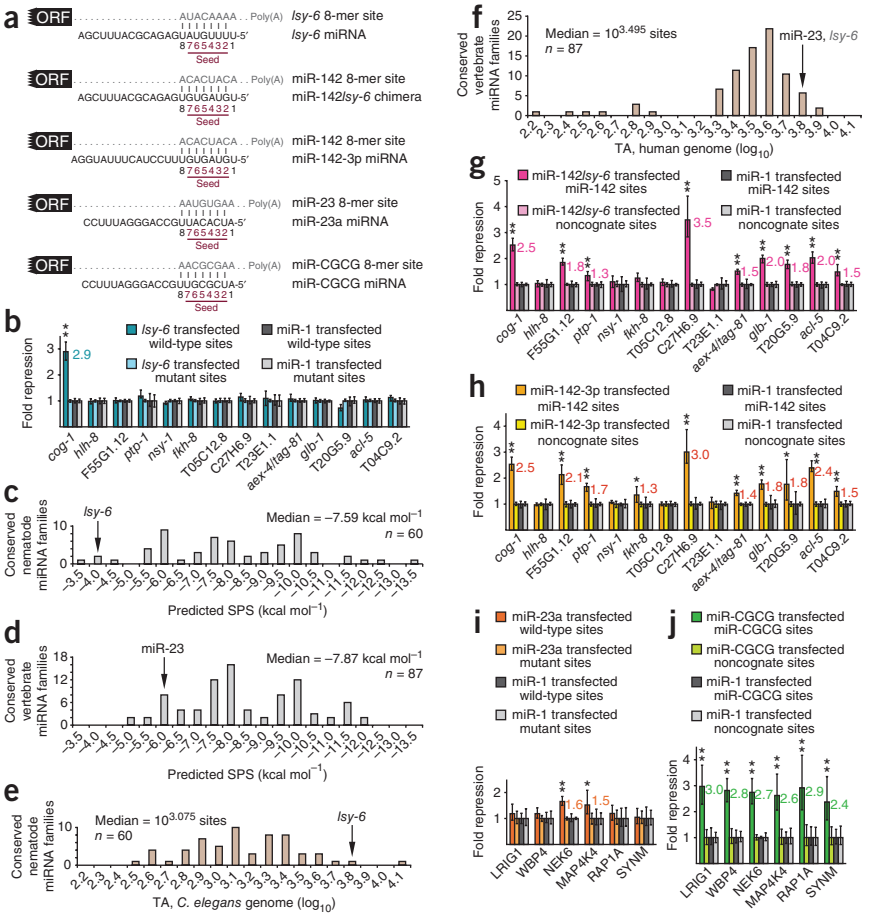

Figure 1 Strengthening SPS while decreasing TA imparted typical targeting proficiency to *lsy-6* and miR-23 miRNAs. (**a**) Sequences of miRNAs and target sites tested in reporter assays. Each miRNA was co-transfected with reporter plasmids as a duplex designed to represent the miRNA paired with its miRNA* strand (**Supplementary Fig. 1a**). (**b**) Response of reporters with 3′ UTRs of predicted *lsy-6* targets after co-transfection with *lsy-6*. As a specificity control, the experiment was also done using a noncognate miRNA, miR-1 (gray bars). Geometric means are plotted relative to those of reporters in which the predicted target sites were mutated after also normalizing for the repression observed for miR-1 (gray bars). Mutant sites of this experiment were the cognate sites of **Figure 2d**. Error bars, third largest and third smallest values among 12 replicates from 4 independent experiments. Significant differences in repression by cognate miRNA compared to that by noncognate miRNA are indicated. (**c**) Distribution of predicted SPSs for 7-mer-m8 sites of 60 conserved nematode miRNA families[36] (**Supplementary Data 2**). Values were rounded down to the next half-integer unit. (**d**) SPS distribution for 7-mer-m8 sites of 87 conserved vertebrate miRNA families[8] (**Supplementary Data 2**). (**e**) Distributions of predicted genome TA for 7-mer-m8 3′ UTR sites of 60 conserved nematode miRNA families (**Supplementary Data 2**). Values were rounded up to the next tenth of a unit. (**f**) Distributions of predicted genome TA for 7-mer-m8 3′ UTR sites of 87 conserved vertebrate miRNA families (**Supplementary Data 2**). (**g**) Response of reporters mutated such that their sites matched the miR-142 seed. The cognate miRNA was the miR-142/*lsy-6* chimera; noncognate sites were *lsy-6* sites. Otherwise, as in **b**. (**h**) As in **g**, except showing the response to miR-142 transfection. (**i**) Response of reporters with 3′ UTRs of predicted miR-23 targets after co-transfection with miR-23a. Noncognate sites were for miR-CGCG. Otherwise, as in **b**. (**j**) Response of reporters mutated such that their sites matched the seed of miR-CGCG, which was co-transfected as the cognate miRNA. Noncognate sites were for miR-23. Otherwise, as in **i**. **P* < 0.01, ***P* < 0.001, Wilcoxon rank-sum test.

8-mer or 7-mer-m8 sites (both 7 base pairs, bp) was weak (−3.65 kcal mol^{-1}), which was weaker than that of all but one conserved nematode miRNA (**Fig. 1c**). The SPS predicted for *lsy-6* was also weaker than that of the weakest of 87 broadly conserved vertebrate miRNAs (**Fig. 1d**). The predicted $\Delta G°$ of an 8-mer or 7-mer-m8 seed match for miR-23 was −5.85 kcal mol^{-1}, in the bottom quintile for broadly conserved vertebrate miRNAs (**Fig. 1d**). We observed similar results for 7-mer-A1 or 6-mer sites (both 6 bp) for both miRNAs (**Supplementary Fig. 1d**,**e**).

lsy-6 is also at the extreme end of the distribution of TA for miRNAs in nematodes and human (**Fig. 1e**,**f**). To predict the TA in a genome, we counted the number of sites in a curated set of distinct 3′ UTRs. When considering a particular cell type, we converted the genome TA to a transcriptome TA by considering the relative levels of each mRNA bearing a site, although in practice the genome and transcriptome TA levels were highly correlated. For example, the transcriptome TA for HeLa cells (TA_{HeLa}) was correlated nearly exactly with the genome TA ($R^2 = 0.98$, $P < 10^{-100}$, Spearman's correlation test, **Supplementary Fig. 1f**). For 8-mer and 7-mer-m8 sites (which both pair with nucleotides 2–8), *lsy-6* had a genome TA that ranked second among 60 *C. elegans* miRNA families and a TA_{HeLa} near that of miR-23, which ranks fifth among the 87 vertebrate families (**Fig. 1e** and **Supplementary Fig. 1g**).

To test the hypothesis that either the weak SPS or high TA of *lsy-6* influences its targeting proficiency, we made three substitutions in the *lsy-6* seed that changed both properties. The three substitutions converted the *lsy-6* seed to that of miR-142-3p (**Fig. 1a** and **Supplementary Fig. 1a**), which changed the predicted SPS to −7.70 kcal mol^{-1}, which was 4.05 kcal mol^{-1} stronger than that of *lsy-6* and near the median values for conserved nematode and vertebrate miRNAs (**Fig. 1c**,**d**). The substitutions also changed the predicted TA to $10^{2.957}$ sites in *C. elegans* and $10^{3.207}$ sites in human, values below the median of conserved miRNAs in both genomes (**Fig. 1e**,**f**). We co-transfected this miR-142*lsy-6* chimeric miRNA and assayed it using reporters with compensatory substitutions in their seed matches, and found it repressed 9 of 14 reporters, a fraction within the range expected in this system using reporters with the site types and contexts assayed (**Fig. 1g**). We repeated the experiment using the full-length miR-142-3p sequence (**Fig. 1a** and **Supplementary Fig. 1a**) and found similar results, indicating that miRNA sequence outside the seed region was irrelevant for repression of both the *cog-1* 3′ UTR and the other *C. elegans* 3′ UTRs (**Fig. 1h**).

Like *lsy-6*, miR-23 also had low targeting proficiency in our system. We surveyed 17 human 3′ UTR fragments, randomly chosen from a set with two 7- to 8-nt miR-23 sites (conserved or nonconserved) spaced within 700 nt of each other, and found that only one fragment was repressed by miR-23 endogenous to either HeLa or HepG2 cells (data not shown). In subsequent experiments focusing on the six UTRs with the most favorable context scores (**Supplementary Table 1**), we found that co-transfecting additional miR-23a imparted marginal or no repression (**Fig. 1i**).

To test whether strengthening SPS while decreasing TA could increase the targeting proficiency of miR-23a, we converted two A:U seed pairs into two G:C pairs (**Fig. 1a** and **Supplementary Fig. 1a**); this strengthened the predicted SPS from −5.85 kcal mol^{-1} to −8.67 kcal mol^{-1} while reducing the TA from the fifth highest of the 87 vertebrate families to below the lowest. We assayed this miRNA, called miR-CGCG, using reporters with compensatory substitutions in their seed matches, and found that the sporadic and marginal repression observed with the wild-type UTRs became much more robust (**Fig. 1j**). These results indicate that miR-23a had low targeting proficiency because of its weak SPS, its high TA, or both, thereby extending our findings to a mammalian miRNA and mammalian 3′ UTRs.

Separating the effects of SPS and TA on miRNA targeting

To differentiate the potential effects of SPS from those of TA, we considered the relationship between these two properties for all 16,384 possible heptamers. In the *C. elegans* 3′ UTRs, these properties were highly anticorrelated (**Fig. 2a**, $R^2 = 0.680$, $P < 10^{-100}$, Spearman's correlation test). In mammalian 3′ UTRs the relationship was still highly significant, but the substantial depletion of CG dinucleotides in the vertebrate transcriptome[31] created more spread in TA, which led to lower correlation coefficients for both human (**Fig. 2b**, $R^2 = 0.121$, $P < 10^{-100}$) and mouse (**Supplementary Fig. 2a**, $R^2 = 0.081$, $P < 10^{-100}$). In general, each additional CG dinucleotide imparted an additional $\log_{10}$ reduction in TA.

To test the influence of TA on *lsy-6* targeting proficiency, we designed the low-TA (LTA) version of *lsy-6*, which had two point substitutions in the *lsy-6* seed (**Fig. 2c** and **Supplementary Fig. 1a**). Substituting U4 with a C (substitution U4C) introduced a CG dinucleotide, whereas the other substitution, U2A, facilitated later investigation of SPS. Because of the CG dinucleotide, LTA-*lsy-6* had a predicted TA_{HeLa} 95% lower than that of *lsy-6*, a value that would be third lowest among the conserved vertebrate miRNA families. Although the substitutions also led to stronger SPS, the predicted SPS of −5.49 kcal mol^{-1} was still slightly weaker than that of miR-23 and well below the median for both nematode and vertebrate conserved miRNAs (**Fig. 1c**,**d**). When assayed using reporters with compensatory substitutions in their seed matches, LTA-*lsy-6* repressed the *cog-1* reporters and only three others (**Fig. 2d**). Two reporters (F55G1.12 and C27H6.9) were repressed only marginally (<1.3 fold), reminiscent of the marginal repression imparted by miR-23 when using its cognate sites. For the third reporter, T20G5.9, we attributed much of the apparent repression to normalization to the miR-1 results, which in the case of this UTR were unusual (**Supplementary Fig. 2d**). Taken together, the LTA-*lsy-6* results indicate that lowering TA was not sufficient alone to confer robust targeting proficiency.

To strengthen SPS without changing TA, we replaced each of the two seed adenines of LTA-*lsy-6* with 2,6-di-aminopurine (DAP or D). DAP is an adenine analog with an exocyclic amino group at position 2, enabling it to pair with uracil with geometry and thermodynamic stability resembling that of a G:C pair (**Fig. 2e**). Because nearest-neighbor parameters had not been determined for model duplexes containing D:U pairs, we estimated SPS by using the values for A:U pairs and adding −0.9 kcal mol^{-1} for each D:U pair, as this is the value of an additional hydrogen bond in model duplexes[32]. With this approximation, the D-LTA-*lsy-6* miRNA had a predicted SPS of −7.29 kcal mol^{-1}, which approached −7.87 kcal mol^{-1}, the median predicted SPS of the conserved vertebrate miRNAs. When assayed using the same reporters as used for LTA-*lsy-6*, D-LTA-*lsy-6* repressed 7 of 14 reporters (**Fig. 2f**). Although this repression was weaker than that observed with the miR-142 seed (**Fig. 1g**,**h**), it was greater than that observed for LTA-*lsy-6* and on par with that expected for mammalian miRNAs in this system using reporters with the site types and site contexts assayed.

We next tested D-miR-23, which also had two seed adenines replaced by DAP, thereby strengthening the predicted SPS from −5.85 kcal mol^{-1} to −7.65 kcal mol^{-1}. Five of the six reporters with miR-23 sites showed significantly greater repression by D-miR-23a than by wild-type miR-23a (**Fig. 2g**), demonstrating a favorable effect for increasing SPS in the context of very high TA (93rd percentile). However, repression was still considerably lower than that conferred by miR-CGCG, presumably because miR-CGCG had lower TA and somewhat stronger SPS (−8.67 kcal mol^{-1}), although we cannot exclude the possibility that the non-natural DAP in the miRNA compromised activity.

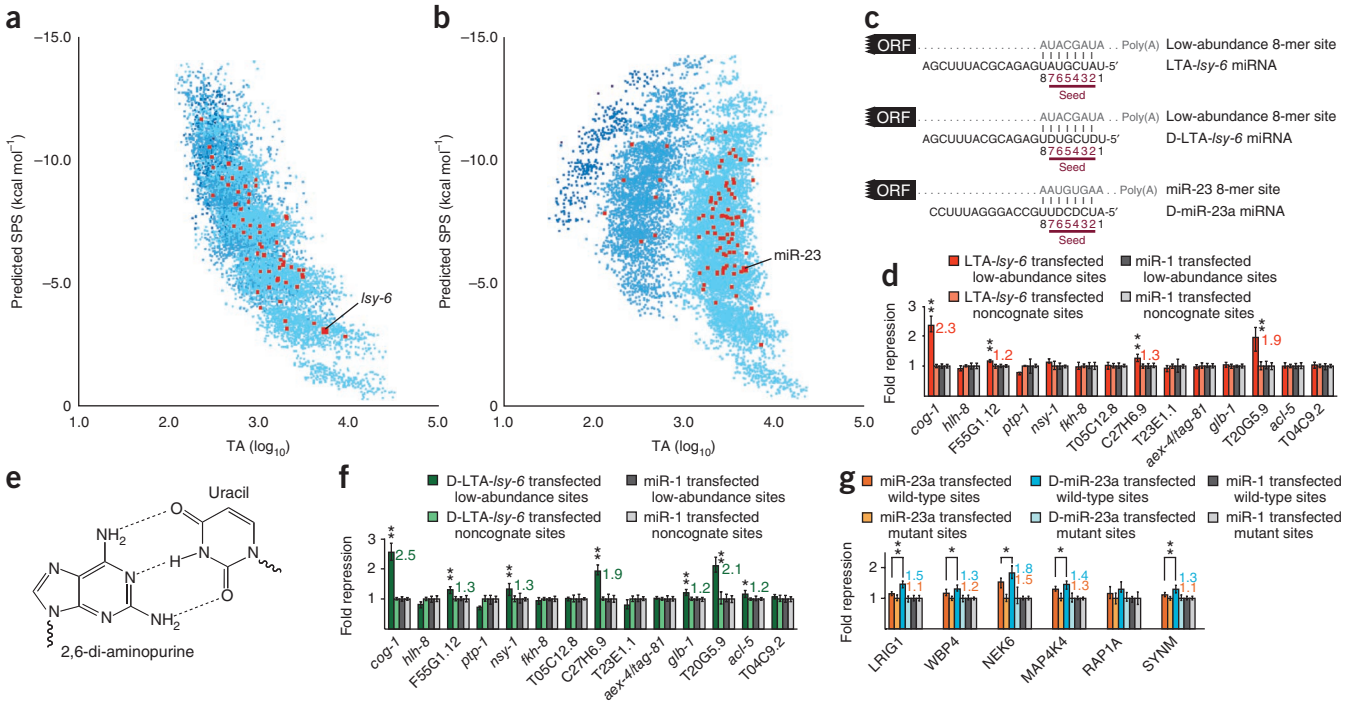

Figure 2 Separating the effects of SPS and TA on miRNA targeting proficiency. (**a**) Relationship between predicted SPS and genomic TA for *lsy-6* and the 59 other conserved nematode miRNAs (red squares), and all other heptamers (light blue, blue, dark blue or purple squares indicating 0, 1, 2 or 3 CpG dinucleotides within the heptamer, respectively). TA was defined as total number of canonical 7- to 8-nt sites (8-mer, 7-mer-m8 and 7-mer-A1) in annotated 3′ UTRs. SPS values were predicted using the respective 7-mer-m8 sites. (**b**) Relationship between predicted SPS and TA in human 3′ UTRs for miR-23 and the 86 other broadly conserved vertebrate miRNA families (red squares). Otherwise, as in **a**. (**c**) Sequences of miRNAs and target sites tested in reporter assays of this figure. (**d**) Response of reporters with 3′ UTRs of predicted *lsy-6* targets mutated such that their sites matched the seed of LTA-*lsy-6*, which was co-transfected as the cognate miRNA. Noncognate sites were for *lsy-6*. Otherwise, as in **Figure 1b**. (**e**) 2,6-di-aminopurine (DAP or D)-uracil base pair. (**f**) Response of reporters used in **d** after co-transfecting D-LTA-*lsy-6* as the cognate miRNA. Otherwise, as in **d**. (**g**) Response of reporters used in **Figure 1i** after co-transfecting D-miR-23a as the cognate miRNA, alongside results for miR-23a that was repeated in parallel. Otherwise, as in **Figure 1i**. $*P < 0.01$, $**P < 0.001$, Wilcoxon rank-sum test.

The results for DAP-substituted miRNAs show that for miRNAs with weak SPS, strengthening SPS can enhance targeting proficiency, regardless of whether these miRNAs have high or low TA. Because DAP substitution changed the predicted SPS without changing the sites in the UTRs, these results indicate that the low proficiency was due to weak SPS rather than occlusion of the sites by RNA-binding proteins that recognized the miRNA seed matches. Taken together, our reporter results also suggest that lowering TA can further enhance targeting proficiency, particularly for miRNAs with moderate to strong SPS.

Global impact of TA and SPS on targeting proficiency

To examine the global impact of TA and SPS on targeting, we collected 175 published microarray data sets that monitored the response of transfecting miRNAs or siRNAs (together referred to as sRNAs) into HeLa cells (**Supplementary Data 1**). Data sets reporting the effects of sRNAs with the same seed region were combined, yielding results for 102 distinct seeds that covered a broad spectrum of TA and predicted SPS (**Fig. 3a**). For each of these 102 data sets, we determined the mean repression of mRNAs with a single 3′ UTR 8-mer site and no other sites in the message, and plotted these values with respect to both the TA_{HeLa} and predicted SPS of the transfected sRNA (**Fig. 3b**, top). sRNAs with lower TA_{HeLa} were more effective than those with higher TA_{HeLa}, and those with stronger predicted SPS were more effective than those with weaker predicted SPS ($P = 0.0006$ and 0.0054 for TA_{HeLa} and SPS, respectively, Pearson's correlation test; **Table 1**). We used multiple linear regression to account for the cross-correlation between TA_{HeLa} and SPS and found that correlations were at least marginally significant for the individual features ($P = 0.005$ and 0.05, *t*-test; **Table 1**), indicating that both properties were independently associated with the proficiency of targeting 3′ UTR sites. We observed similar results for targeting 7-mer-m8, 7-mer-A1 and 6-mer sites (**Fig. 3b** and **Table 1**).

Although both TA and SPS each significantly influenced targeting proficiency, together they explained only a minority of the variability (**Table 1**). Most of the variability could be from factors unrelated to targeting, such as array noise, differential transfection efficiencies or differential sRNA loading or stability. To reduce variability from these sources, we focused on 74 data sets for which responsive messages were significantly enriched in 3′ UTR sites to the transfected sRNA (**Fig. 3a**, red squares; **Supplementary Data 1**). In these filtered data sets, correlations between proficiency and both TA_{HeLa} and SPS were stronger and observed with similar significance, even though the filtering reduced the quantity of data analyzed and might have preferentially discarded data sets for which high TA or weak SPS prevented detectable repression (**Supplementary Fig. 3a**,**b** and **Supplementary Table 2**).

Studies monitoring global effects of miRNAs on target repression have concluded that sites in open reading frames (ORFs) can mediate repression but that the efficacy of these sites is generally less than that of sites in 3′ UTRs[7,30,33,34]. To examine the impact of TA and SPS on targeting in ORFs, we considered expressed messages that had a single

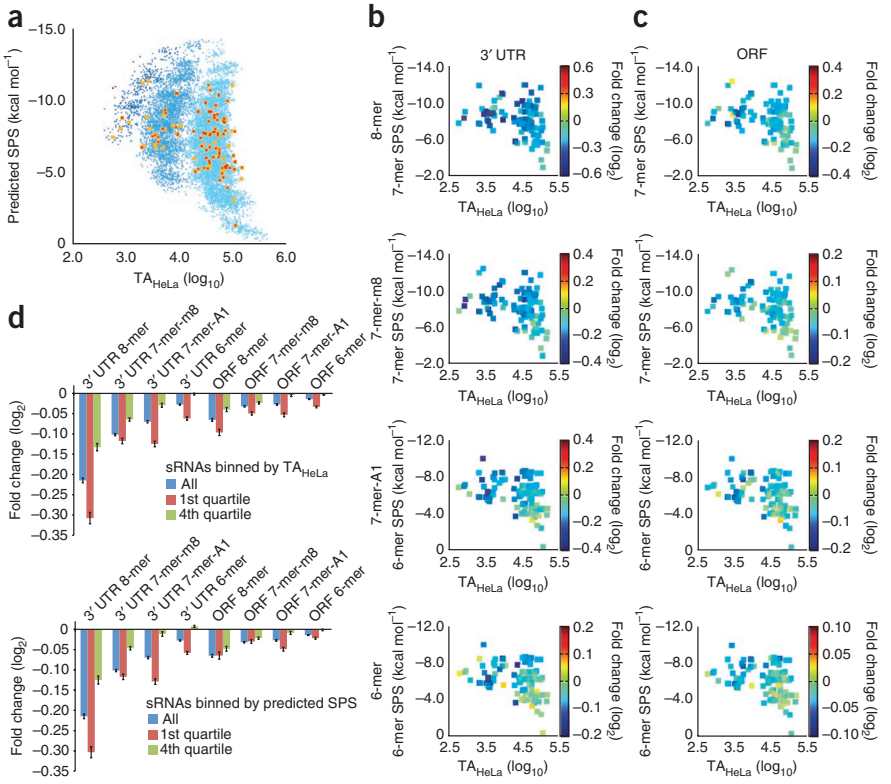

Figure 3 Impact of TA and SPS on sRNA targeting proficiency, as determined using array data. (**a**) Distribution of TA_{HeLa} and predicted SPS for the sRNAs from the 102 array data sets analyzed in this study (orange squares) and sRNAs from data sets that passed the motif-enrichment analysis (red squares). Otherwise, plotted as in **Figure 2b**. (**b**) Response of expressed mRNAs with a single 3′ UTR site to the cognate sRNA, with respect to TA_{HeLa} and predicted SPS. Fold-change values are plotted according the key to the right of each plot, comparing mRNAs with a single site of the type indicated (and no additional sites to the cognate sRNA elsewhere in the mRNA) to those with no site to the cognate sRNA; note different scales for different plots. In areas of overlap, mean values are plotted. Correlation coefficients and *P* values are in **Table 1**. (**c**) Response of expressed mRNAs with a single ORF site to the cognate sRNA, with respect to TA_{HeLa} and predicted SPS. Otherwise, as in **b**. (**d**) Response of mRNAs with indicated single sites when binning cognate sRNA by TA_{HeLa} (top) or predicted SPS (bottom). The key indicates the data considered, with the first quartiles at top comprising data for sRNAs with the lowest TA_{HeLa} and those at bottom comprising data for sRNAs with the strongest predicted SPS. Error bars, 95% confidence intervals.

ORF site but no additional sites in the rest of the message. For 7-mer-m8 and 6-mer sites, mean repression was significantly correlated with both TA_{HeLa} and predicted SPS, and for the other two sites in ORFs, mean repression was significantly correlated with TA_{HeLa} (**Fig. 3c** and **Table 1**). The response of sites in 5′ UTRs was not significantly correlated with either TA or predicted SPS (**Table 1**), consistent with the idea that 5′ UTRs harbor few effective sites[3].

We next examined the quantitative impact of TA and SPS on targeting proficiency. We considered the same sets of mRNAs with single sites to the cognate sRNAs, and for each site type and each mRNA region, we binned mRNAs into quartiles ranked by either low TA or strong predicted SPS. For each site type, messages in the top quartile responded more strongly than those in the bottom (**Fig. 3d**). The differences usually were substantial. For example, repression of the top quartile of mRNAs with 7-mer-A1 sites matched the mean repression of mRNAs with 7-mer-m8 sites, whereas repression of the bottom quartile resembled the mean repression of mRNAs with 6-mer sites.

Improved miRNA target prediction

An effective tool for mammalian miRNA target prediction is the context score[30]. Context scores are used to rank mammalian miRNA target predictions by modeling the relative contributions of previously identified targeting features, including site type, site number, site location, local A+U content and 3′-supplementary pairing, to predict the relative repression of mRNAs with 3′ UTR sites[7]. However, the context-score model was not designed to consider differences between sRNAs, such as TA or SPS, which can cause sites of one miRNA to be more robustly targeted than those of another (assuming equal expression of the two miRNAs).

To build a model appropriate for predicting the relative response of targets of different miRNAs, we considered TA and SPS as two independent variables when carrying out multiple linear regression on the 11 microarray data sets used previously for the initial development and training of the context-score model[7]. The other parameters were local A+U content, the location of the site within the 3′ UTR, and 3′-supplementary pairing[7]. For each site type, TA and/or SPS robustly contributed (**Supplementary Table 3**). The scores generated by these models were called context+ scores, because they consider site type and context plus sRNA proficiency. We then generated the total context+ score for each mRNA with 3′

Table 1 Relationship between mean mRNA repression and either TA or predicted SPS for indicated site types

	Multiple linear regression			Simple linear regression			
		P value		TA_{HeLa}		SPS	
Site location and type	Multiple R^2	TA_{HeLa}	SPS	R^2	*P* value	R^2	*P* value
3′ UTR 8-mer	0.149	0.0049	0.051	0.115	0.0006	0.076	0.0054
3′ UTR 7-mer-m8	0.190	0.0081	0.0047	0.122	0.0003	0.131	0.0002
3′ UTR 7-mer-A1	0.335	0.0009	2×10^{-5}	0.196	3×10^{-6}	0.256	6×10^{-8}
3′ UTR 6-mer	0.177	0.039	0.0025	0.097	0.0014	0.141	0.0001
ORF 8-mer	0.104	0.018	0.14	0.085	0.0030	0.052	0.021
ORF 7-mer-m8	0.171	0.019	0.0054	0.103	0.0010	0.123	0.0003
ORF 7-mer-A1	0.135	0.010	0.073	0.106	0.0008	0.076	0.0052
ORF 6-mer	0.228	0.010	0.0008	0.133	0.0002	0.174	1×10^{-5}
5′ UTR 8-mer	0.004	0.75	0.68	0.002	0.64	0.003	0.59
5′ UTR 7-mer-m8	0.003	0.63	0.72	0.002	0.70	0.000	0.84
5′ UTR 7-mer-A1	0.012	0.60	0.49	0.007	0.41	0.009	0.35
5′ UTR 6-mer	0.011	0.97	0.32	0.001	0.74	0.011	0.29

Results were determined from microarray data (**Fig. 3b,c**).

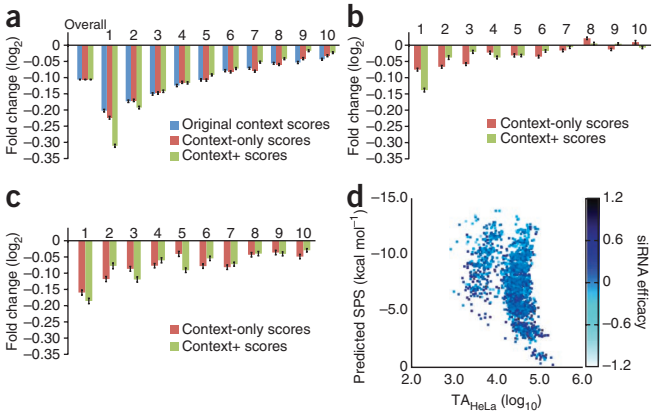

Figure 4 Predictive performance of the context+ model, which considers miRNA or siRNA proficiency in addition to site context. (**a**) Improved predictions for mRNAs with canonical 7- to 8-nt 3′ UTR sites. Predicted interactions between mRNAs and cognate sRNA were distributed into ten equally populated bins based on total context scores generated using the model indicated (key), with the first bin comprising interactions with most favorable scores. Plotted for each bin is the mean mRNA change on the arrays (error bars, 95% confidence intervals). (**b**) Prediction of responsive interactions involving mRNAs with only 3′ UTR 6-mer sites. Otherwise, as in **a**. (**c**) Prediction of responsive interactions involving mRNAs with at least one 8-mer ORF site but no 3′ UTR sites. Otherwise, as in **a**. (**d**) Impact of TA and SPS on siRNA-directed knockdown of the desired target. Efficacy in luciferase activity knockdown for 2,431 siRNAs transfected into H1299 cells[38]. Efficacy is linearly scaled (key), with positive and negative controls having values of 0.900 and 0.354, respectively[38].

UTR sites, relying on the observation that multiple sites typically act independently with respect to each other[7].

We tested the predictive value of the new model using data from array data sets not used to train the model, and comparing the performance of the predicted targets ranked using the total context+ scores to those ranked using scores of the original model. To examine whether any improvement over the original model was due to training the model with multiple linear regression rather than simple linear regression, we also used multiple linear regression to build a model that considered only the three parameters used to build the original model (context-only scores, **Supplementary Table 4**). For each model, we ranked predicted targets with 7- to 8-nt sites by score and assigned them to ten bins. The context+ scores performed better than the old context scores at predicting the response to the sRNAs (**Fig. 4a**), yielding significantly stronger mean repression for the top two bins ($P = 5 \times 10^{-56}$ and 3×10^{-8} for bins 1 and 2, respectively) and significantly weaker repression in the bottom four bins ($P = 6 \times 10^{-10}$, 1.5×10^{-5}, 1×10^{-7} and 3×10^{-4} for bins 7–10, respectively, Wilcoxon's rank sum test). Improved specificity was also demonstrated in receiver operating characteristic (ROC) curves (**Supplementary Fig. 4a**).

Because most 6-mer sites and ORF sites are either nonresponsive or only marginally responsive to the miRNA, algorithms that achieve useful prediction specificity do so at the expense of ignoring these sites[3]. As low TA and strong SPS were correlated with substantially greater efficacy of these marginal sites (**Fig. 3c**,**d**), we extended the context+ scores to 6-mer sites. For the context+ model, the top bin of mRNAs with 6-mer 3′ UTR sites but no larger sites (**Fig. 4b**) had average repression resembling that of the third bin of mRNAs with 7- to 8-nt 3′ UTR sites (**Fig. 4a**; ROC curves, **Supplementary Fig. 4b**). We also generated context-only and context+ scores for ORF sites by changing only the parameter of site location; this was not applicable for ORF sites because it accounts for the lower efficacy of sites near the middle of long 3′ UTRs[7]. In ORFs, we found that sites farther from the stop codon tended to be less effective, and thus we included the distance from the stop codon (linearly scaled distance of 0 to ≥1,500 nt) as a parameter. Although this context+ model was not substantially more predictive than the context-only model for ORF sites (perhaps because data from only 11 miRNAs were used in the regression), both models had predictive value. We compared mRNAs with at least one 8-mer ORF site (**Fig. 4c**) and found that those ranked in the top bin had average repression resembling that of the second or third bins of mRNAs with 7- to 8-nt 3′ UTR sites (**Fig. 4a**).

Overall, our findings show that taking TA and SPS into account can significantly improve miRNA target prediction when pooling results from multiple sRNAs. Training on the 11 miRNA transfection data sets used for the original context scores was appropriate for demonstrating the improvement that could be achieved by taking TA and SPS into account. We reasoned, however, that training on the 74 filtered data sets could generate a more precise context+ model to be used to quantitatively predict repression. As we expected, correlations for all four parameters had even greater significance when we trained the model on more data (**Supplementary Table 5**). Although a support vector machine (SVM) approach should in principle yield even greater specificity by capturing effects lost in multiple linear regression due to multicollinearity, we did not observe enhanced performance with SVM (**Supplementary Fig. 4c–e**). Therefore, we used multiple linear regression because it enabled more convenient calculation of context+ scores (**Supplementary Fig. 5a**). We will use these new scores in version 6.0 of TargetScan (http://www.targetscan.org/).

Additional considerations

A caveat of the reporter experiments was that miRNA sequence changes designed to alter TA or SPS could have influenced other factors, such as miRNA stability or its loading into the silencing complex. However, our computational analyses of 102 array data sets also showed that TA and SPS each independently influence targeting efficacy. Therefore, if differences in sRNA stability or loading confounded interpretation of our results, these differences would be correlated with either predicted SPS or TA. Analysis of published miRNA overexpression data countered this possibility, showing no correlation between miRNA accumulation and predicted SPS or TA (**Supplementary Fig. 3c**,**d**). Furthermore, experiments examining the RNAs co-purifying with AGO2 indicated that the difference in proficiency between *lsy-6* and miR-142*lsy-6* was not merely attributable to less accumulation of *lsy-6* in the silencing complex (**Supplementary Fig. 1m–s**).

DISCUSSION

The correlation between strong SPS and low TA has confounded earlier efforts to examine the influence of these parameters on targeting efficacy, with one study implicating SPS and not TA[24] and others implicating TA and not SPS[27,28]. Our results indicate that both parameters influence efficacy and solve one of the mysteries in miRNA targeting, the failure of *lsy-6* to repress all but one of the 14 examined seed-matched mRNAs. Previous studies have hypothesized that the seed-based targeting model is unreliable[23] or that sites of the 13 nonresponsive mRNAs fall in inaccessible UTR structure[18]. Our work shows that the solution is the unusually weak SPS and high TA of the *lsy-6* miRNA. Changing these parameters to resemble those of more typical miRNAs imparted typical seed-based targeting proficiency, even though the sites were in their original UTR contexts, thereby

demonstrating that neither the reliability of seed-based targeting nor the accessibility of the sites were at issue.

MicroRNAs with unusually weak predicted SPS and unusually high TA, such as miR-23 and *lsy-6*, seem to have few targets. Indeed, *lsy-6* might have only a single biological target, the *cog-1* mRNA—an extreme exception to the finding that metazoan miRNAs generally have dozens if not hundreds of preferentially conserved targets[4,8,35,36]. Determining why so few mRNAs respond to *lsy-6* brings to the fore a second mystery, still unsolved: how is the *cog-1* 3′ UTR so efficiently recognized and repressed by a miRNA with such weak targeting proficiency? This UTR has two 8-mer sites, which by virtue of their conservation make *cog-1* the top predicted target of *lsy-6* (ref. 3), but this is only part of the answer[37]. Improving the context-score model to take into account the differential SPS and TA of different miRNAs may help focus attention on the predicted targets of miRNAs with more typical proficiencies, but leaves unsolved the problem of how to predict the few biological sites of the less proficient miRNAs without considering site conservation.

MicroRNAs with very high TA, such as *lsy-6* or miR-23, and those with very low TA, such as miR-100 or miR-126, two broadly conserved vertebrate miRNAs containing CG dinucleotides in their seeds (**Supplementary Data 2**), seem to represent two strategies for targeting very few genes, accomplished at opposite ends of the TA spectrum. For miRNAs with very high TA, other UTR features flanking the seed sites are required for regulation, as has been shown for *lsy-6* regulation of *cog-1* (ref. 37), whereas miRNAs with very low TA have far fewer potential target sites to begin with.

Our results also have implications for how siRNA could be designed to reduce off-targets. Earlier studies have proposed that off-targets could be reduced by designing siRNAs with low TA[27] or weak SPS[24], and our results suggest that off-targets could be largely eliminated by designing siRNAs with both high TA and weak SPS. However, such siRNAs might also be ineffective at recognizing the desired mRNA target because pairing with this target would nucleate on a match with weak SPS and might be titrated by the many other mRNAs with seed matches. To investigate this concern, we examined a published data set of high-throughput luciferase assays reporting the response to 2,431 different siRNAs[38]. siRNAs with weak predicted SPS knocked down the desired target more effectively than did those with strong predicted SPS (**Fig. 4d**; $P < 10^{-100}$, *t*-test), presumably because of preferential loading into the silencing complex[39,40]. Moreover, high TA did not compromise the desired targeting efficacy, even after we corrected for the cross-correlation between TA and SPS ($P = 0.16$, *t*-test). Therefore, designing siRNAs with high TA and weak SPS should minimize off-target effects without compromising knockdown of the desired target.

Highly expressed mRNAs tend to be evolutionarily depleted in sites for coexpressed miRNAs, a phenomenon partly attributed to the possibility that these mRNAs might otherwise titrate the miRNAs from their intended targets[12,41,42]. Titration can also provide a useful mechanism for cells to regulate miRNA activity, as has been shown by *IPS1* titration of miR-399 in *Arabidopsis thaliana*[25]. Beneficial titration has even been proposed to explain why so many miRNA sites are conserved[43]. However, because most preferentially conserved sites are in lowly to moderately expressed mRNAs, and because these sites each comprise only a tiny fraction of the TA, each could impart at most a correspondingly tiny effect on the effective miRNA concentration—much less than that required to selectively retain the site. Although titration functions cannot explain most site conservation, TA could be dynamic during development, with notable consequences. For example, the increase of a miRNA during development is often accompanied by a decrease in its transcriptome TA, a consequence of the evolutionary depletion of sites in mRNAs coexpressed at high levels with the miRNA[12,42]. This accompanying TA decrease would sharpen the transition between the nonrepressed and repressed states of targets.

When predicting SPS, we used parameters derived from model RNA duplexes, which presumably underestimated the affinity of RNA segments pairing with Argonaute-bound seed regions[2,3,44,45]. The extent to which Argonaute enhances affinity might vary for different seed sequences. These potential differences, however, did not obscure our detection of an influence of SPS on targeting proficiency. Thus, our study provides a lower bound on the influence of SPS, and an approach for determining its full magnitude once accurate SPSs of Argonaute-bound complexes are known.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

ACKNOWLEDGMENTS

We thank D. Didiano and O. Hobert (Columbia University) for *lsy-6* target constructs and V. Auyeung, R. Friedman, C. Jan and H. Guo for helpful discussions and for sharing data sets before publication. This work was supported by US National Institutes of Health grant GM067031 (D.P.B.) and a Research Settlement Fund for the new faculty of SNU (D.B.). D.P.B. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

D.M.G. carried out most reporter assays and associated experiments and analyses. D.B. carried out all the computational analyses except for reporter analyses. G.W.B. implemented revisions to the TargetScan site. C.S. and A.G. carried out assays and analyses involving miR-23. D.M.G., D.B. and D.P.B. wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/nsmb/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
2. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
3. Bartel, D.P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
4. Lewis, B.P., Burge, C.B. & Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
5. Brennecke, J., Stark, A., Russell, R.B. & Cohen, S.M. Principles of microRNA-target recognition. *PLoS Biol.* **3**, e85 (2005).
6. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat. Genet.* **37**, 495–500 (2005).
7. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27**, 91–105 (2007).
8. Friedman, R.C., Farh, K.K., Burge, C.B. & Bartel, D.P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19**, 92–105 (2009).
9. Shin, C. *et al.* Expanding the microRNA targeting code: functional sites with centered pairing. *Mol. Cell* **38**, 789–802 (2010).
10. Lim, L.P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
11. Krützfeldt, J. *et al.* Silencing of microRNAs *in vivo* with 'antagomirs'. *Nature* **438**, 685–689 (2005).
12. Farh, K.K. *et al.* The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science* **310**, 1817–1821 (2005).
13. Giraldez, A.J. *et al.* Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**, 75–79 (2006).
14. Nielsen, C.B. *et al.* Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* **13**, 1894–1910 (2007).

15. Robins, H., Li, Y. & Padgett, R.W. Incorporating structure to predict microRNA targets. *Proc. Natl. Acad. Sci. USA* **102**, 4006–4009 (2005).
16. Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets *Hand2* during cardiogenesis. *Nature* **436**, 214–220 (2005).
17. Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U. & Segal, E. The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39**, 1278–1284 (2007).
18. Long, D. *et al.* Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.* **14**, 287–294 (2007).
19. Hammell, M. *et al.* mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein-enriched transcripts. *Nat. Methods* **5**, 813–819 (2008).
20. Saetrom, P. *et al.* Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res.* **35**, 2333–2342 (2007).
21. Kedde, M. *et al.* RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell* **131**, 1273–1286 (2007).
22. Ruby, J.G. *et al.* Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res.* **17**, 1850–1864 (2007).
23. Didiano, D. & Hobert, O. Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat. Struct. Mol. Biol.* **13**, 849–851 (2006).
24. Ui-Tei, K., Naito, Y., Nishi, K., Juni, A. & Saigo, K. Thermodynamic stability and Watson-Crick base pairing in the seed duplex are major determinants of the efficiency of the siRNA-based off-target effect. *Nucleic Acids Res.* **36**, 7100–7109 (2008).
25. Franco-Zorrilla, J.M. *et al.* Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.* **39**, 1033–1037 (2007).
26. Ebert, M.S., Neilson, J.R. & Sharp, P.A. MicroRNA sponges: competitive inhibitors of small RNAs in mammalian cells. *Nat. Methods* **4**, 721–726 (2007).
27. Anderson, E.M. *et al.* Experimental validation of the importance of seed complement frequency to siRNA specificity. *RNA* **14**, 853–861 (2008).
28. Arvey, A., Larsson, E., Sander, C., Leslie, C.S. & Marks, D.S. Target mRNA abundance dilutes microRNA and siRNA activity. *Mol. Syst. Biol.* **6**, 363 (2010).
29. Rodriguez, A. *et al.* Requirement of bic/microRNA-155 for normal immune function. *Science* **316**, 608–611 (2007).
30. Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008).
31. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
32. Xia, T. *et al.* Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* **37**, 14719–14735 (1998).
33. Guo, H., Ingolia, N.T., Weissman, J.S. & Bartel, D.P. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**, 835–840 (2010).
34. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).
35. Lall, S. *et al.* A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr. Biol.* **16**, 460–471 (2006).
36. Jan, C.H., Friedman, R.C., Ruby, J.G. & Bartel, D.P. Formation, regulation and evolution of *Caenorhabditis elegans* 3′ UTRs. *Nature* **469**, 97–101 (2011).
37. Didiano, D. & Hobert, O. Molecular architecture of a miRNA-regulated 3′ UTR. *RNA* **14**, 1297–1317 (2008).
38. Huesken, D. *et al.* Design of a genome-wide siRNA library using an artificial neural network. *Nat. Biotechnol.* **23**, 995–1001 (2005).
39. Schwarz, D.S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199–208 (2003).
40. Khvorova, A., Reynolds, A. & Jayasena, S.D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209–216 (2003).
41. Bartel, D.P. & Chen, C.Z. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat. Rev. Genet.* **5**, 396–400 (2004).
42. Stark, A., Brennecke, J., Bushati, N., Russell, R.B. & Cohen, S.M. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123**, 1133–1146 (2005).
43. Seitz, H. Redefining microRNA targets. *Curr. Biol.* **19**, 870–873 (2009).
44. Ameres, S.L., Martinez, J. & Schroeder, R. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell* **130**, 101–112 (2007).
45. Parker, J.S., Parizotto, E.A., Wang, M., Roe, S.M. & Barford, D. Enhancement of the seed-target recognition step in RNA silencing by a PIWI/MID domain protein. *Mol. Cell* **33**, 204–214 (2009).

ONLINE METHODS

Reporter assays. For *lsy-6* reporter assays, HeLa cells were plated in 24-well plates at 5×10^4 cells per well. After 24 h, each well was transfected with 20 ng TK-*Renilla*-luciferase reporter (pIS1)[46], 20 ng firefly-luciferase control reporter (pIS0)[46] and 25 nM miRNA duplex (Dharmacon; **Supplementary Fig. 1a**), using Lipofectamine 2000 (Invitrogen). For miR-23 reporter assays, conditions were the same except for transfected DNA: 10 ng SV40-*Renilla*-luciferase reporter (pIS2)[46], 25 ng firefly-luciferase control reporter (pIS0) and 1.25 µg pUC19 carrier DNA. Luciferase activities were measured 24 h after transfection with the Dual-Luciferase Assay (Promega) and a Veritas microplate luminometer (Turner BioSystems). For every construct assayed, four independent experiments, each with three biological replicates, were done. To control for transfection efficiency, firefly activity was divided by *Renilla* activity. Values for constructs with sites matching the cognate miRNA were then normalized to the geometric mean of values for otherwise identical constructs in which the sites were mutated. To control for differences not attributable to the cognate miRNA, the ratios were further normalized to ratios for the same constructs tested with a noncognate miRNA, miR-1. These double-normalized results are in figures; singly normalized results are in **Supplementary Figures 1h–l** and **2d–f**.

Constructs. 3′ UTRs of *lsy-6* predicted targets[23] were subcloned into XbaI and EagI sites in pIS1, and 3′ UTRs of miR-23 predicated targets were cloned into SacI and SpeI sites in pIS2 after amplification (UTR sequences, **Supplementary Table 1**). Mutations were introduced using Quikchange (Stratagene) and confirmed by sequencing.

Predicted SPS. SPS was predicted using nearest-neighbor thermodynamic parameters, including the penalty for terminal A:U pairs[32]. The contribution of the A at position 1 of 8-mer and 7-mer-A1 sites was not included because this A does not pair with the miRNA[4] and thus its contribution is not expected to differ predictably for different miRNAs. For linear regression analyses, the predicted SPS of positions 2–8 was used for 8-mer and 7-mer-m8 sites, and the predicted SPS of positions 2–7 was used for 7-mer-A1 and 6-mer sites. To assign a single value for 7- to 8-nt sites (7-mer-A1, 7-mer-m8 and 8-mer), we used a mean weighted value of the three site types. This mean SPS was calculated as [(6-mer SPS)(7-mer-A1 TA) + (7-mer-m8 SPS)(7-mer-m8 TA + 8-mer TA)] / (7-mer-A1 TA + 7-mer-m8 TA + 8-mer TA).

Reference mRNAs. To generate a list of unique mRNAs, human full-length mRNAs obtained from RefSeq[47] and H-Invitational[48] databases were aligned to the human genome[49] (hg18) using BLAT[50] software and processed as described to represent each gene by the mRNA isoform with the longest UTR[30]. These unique full-length mRNAs, which were each represented by the genomic sequence of their exons (as the genomic sequence was of higher quality than the mRNA sequence), were the reference mRNAs (**Supplementary Data 3**). Mouse full-length mRNAs were obtained from RefSeq[47] and FANTOM DB[51] databases, aligned against the mouse genome[52] (mm9) and processed similarly. For *C. elegans* and *Drosophila melanogaster*, we obtained 3′ UTR sequences from TargetScan (targetscan.org)[22,53]. Mature miRNA sequences were downloaded from the miRBase web site[54].

Microarray processing and mapping to reference mRNAs. We collected published data sets reporting the response of HeLa mRNAs 24 h after 100 nM sRNA transfection using Agilent arrays (two-color platform), excluding data sets for which either multiple sRNAs were simultaneously transfected or the transfected RNAs contained chemically modified nucleotides (**Supplementary Data 1**). If probe sequences for an array platform were available, they were mapped to genomic locations in the human genome using BLAT[50] software. For some arrays (for example, GSE8501), probe sequences were unavailable, but associated cDNA or EST sequence IDs were available. In such cases, genomic coordinates of cDNAs and ESTs obtained from the UCSC Genome Browser[55] were used as if they were coordinates of array probes. Each probe and its associated mRNA fold-change value were mapped to the reference mRNA sharing the greatest overlap with the probe's genomic coordinates, ≥15 bases. When multiple probes were mapped to a single reference mRNA, the median fold change was used. To avoid analysis of mRNAs not expressed in HeLa cells, only mRNAs with signal greater than the median in the mock-transfection samples were considered. For each array, the median fold change of reference mRNAs without any 6- to 8-nt site was used to normalize the fold changes of all reference mRNAs. To correct for the global association between mRNA fold change and A+U content of the mRNA transcript, the LOWESS filtering was applied by using the malowess function within MATLAB (Mathworks) (**Supplementary Data 4**). For some arrays, the transfected sRNA was designed to target nearly perfectly matching (≥18 nt) mRNAs, in which case these intended targets were excluded from analysis.

Motif-enrichment analysis for array filtering. To evaluate array data sets, we carried out motif-enrichment analysis using the Fisher's exact test for a 2×2 contingency table, populated based on whether the reference mRNA had a 7-mer motif for the cognate sRNA in its 3′ UTR and whether it was among the top 5% most downregulated mRNAs. If multiple arrays examined the effects of transfecting sRNAs with identical seed regions (positions 2–8), the *P* value of the Fisher's exact test for site enrichment (considering either of the two 7-mer sites and picking the one with the lower *P* value) was assessed for each array, and the array with the median *P* value was chosen to represent that seed region, yielding 102 representative arrays (**Supplementary Data 1**). To obtain a filtered data set, this test was repeated for the 16,384 heptamers, and arrays were retained if the motif most significantly associated with downregulation was the 7-mer-m8 or 7-mer-A1 site of the transfected sRNA; 74 arrays passed this filter (**Supplementary Data 1**). Results of multiple linear regression and other analyses were robust to cutoff choice (other cutoffs tested were 10, 15 and 20%; data not shown).

Target site abundance. TA in the human transcriptome was calculated as the number of nonoverlapping 3′ UTR 8-mer, 7-mer-m8 and 7-mer-A1 sites in the reference mRNAs. An analogous process was used to calculate TA in mouse, *C. elegans* and *D. melanogaster*. To calculate TA_{HeLa}, each site was weighted based on mRNA-Seq data[33]. Predicted SPS and TA values for all heptamers in *C. elegans*, human and HeLa, mouse and *D. melanogaster* are in **Supplementary Data 5**.

miRNA target prediction and analysis of siRNA efficacy. Context scores were calculated for the cognate sites of the reference mRNAs using the simple linear regression parameters reported earlier[7]. Before fitting, scores for each parameter were scaled from 0 to 1 (**Supplementary Fig. 5b**). To account for site type without the complication of multiple sites, we developed models for each type individually, using mRNAs with only a single site to the cognate miRNA (**Supplementary Fig. 5c**). The multiple linear regression models for context-only and context+ were computed by using the lm function in the R package version 2.11.1.

46. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. & Burge, C.B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).
47. Pruitt, K.D., Katz, K.S., Sicotte, H. & Maglott, D.R. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet.* **16**, 44–47 (2000).
48. Imanishi, T. *et al.* Integrative annotation of 21,037 human genes validated by full-length cDNA clones. *PLoS Biol.* **2**, e162 (2004).
49. Lander, E.S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, I860–921 (2001).
50. Kent, W.J. BLAT–the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
51. Okazaki, Y. *et al.* Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563–573 (2002).
52. Waterston, R.H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
53. Ruby, J.G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–1207 (2006).
54. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–D158 (2008).
55. Rhead, B. *et al.* The UCSC Genome Browser database: update 2010. *Nucleic Acids Res.* **38**, D613–D619 (2010).