mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein–enriched transcripts

Molly Hammell[1], Dang Long[2], Liang Zhang[3], Andrew Lee[1], C Steven Carmack[2], Min Han[3], Ye Ding[2] & Victor Ambros[1]

Target prediction for animal microRNAs (miRNAs) has been hindered by the small number of verified targets available to evaluate the accuracy of predicted miRNA-target interactions. Recently, a dataset of 3,404 miRNA-associated mRNA transcripts was identified by immunoprecipitation of the RNA-induced silencing complex components AIN-1 and AIN-2. Our analysis of this AIN-IP dataset revealed enrichment for defining characteristics of functional miRNA-target interactions, including structural accessibility of target sequences, total free energy of miRNA-target hybridization and topology of base-pairing to the 5′ seed region of the miRNA. We used these enriched characteristics as the basis for a quantitative miRNA target prediction method, miRNA targets by weighting immunoprecipitation-enriched parameters (mirWIP), which optimizes sensitivity to verified miRNA-target interactions and specificity to the AIN-IP dataset. MirWIP can be used to capture all known conserved miRNA-mRNA target relationships in *Caenorhabditis elegans* at a lower false-positive rate than can the current standard methods.

The discovery of miRNAs[1] and their roles in post-transcriptional gene regulation has added a new dimension to the study of animal development and disease[2]. miRNAs, bound to their mRNA targets, can repress gene expression through translational inhibition or by mRNA destabilization[3]. Under some conditions, miRNAs may also promote protein production from a target mRNA[4]. Animal miRNAs have a role in regulating many developmental processes and have been implicated in human disease pathways[5]. For these reasons, it is crucial to efficiently identify the functionally important mRNA targets of miRNAs.

Target prediction for miRNAs in plants is straightforward, as plant miRNAs bind with near-perfect complementarity to their target mRNAs. In animals, miRNAs interact with their targets predominantly by partial base-pairing, and the rules that govern the formation and functional efficacy of miRNA-mRNA interactions are not fully understood. Depending on the computational algorithm applied, the number of predicted targets for a given miRNA can range from dozens to hundreds and even thousands of genes[6,7]. The thorough experimental testing of such vast numbers of predicted targets using labor-intensive transgenic reporter assays is impractical. There remains the need both for more accurate computational methods to identify functional miRNA-target interactions and for more efficient methods to experimentally validate miRNA-target interactions *in vivo*.

Many computational methods have been developed to predict miRNA targets (reviewed in ref. 7). The criteria for target prediction vary widely, but often include (i) strong Watson-Crick base-pairing of the 5′ seed of the miRNA (nucleotide positions 2–8 of the miRNA) to a complementary site in the 3′ untranslated region (UTR) of the mRNA, (ii) conservation of the miRNA binding site, (iii) favorable minimum free energy (MFE) for the local miRNA-mRNA interaction, and/or (iv) structural accessibility of the surrounding mRNA sequence. Experimental support exists for each of these binding-site features, but the relative importance of each feature and how they interact to contribute to function remains uncertain. Moreover, it is likely that other important parameters for functional miRNA-target interactions remain to be identified.

The principle of 5′ seed primacy in miRNA-target binding is well supported by experimental data. Many genetically validated miRNA-target interactions involve uninterrupted Watson-Crick base-pairing in the 5′ region of the miRNA. Experiments show that G-U wobble pairs and bulges within this seed region can significantly disrupt repression of reporter constructs[8] and that perfectly matched seed regions are significantly enriched in the 3′ UTRs of transcripts whose levels decrease in response to miRNA overexpression[9]. However, other experimental data suggest that perfectly matched miRNA seeds are neither necessary nor sufficient for all functional miRNA-target interactions. For instance, three of the genetically verified *let-7* targets in *C. elegans*—*lin-41, pha-4* and *let-60*—contain only imperfect binding sites, with G-U wobble

[1]Program in Molecular Medicine, University of Massachusetts Medical School, 373 Plantation Street, Suite 306, Worcester, Massachusetts 01605, USA. [2]Wadsworth Center, New York State Department of Health, 150 New Scotland Avenue, Albany, New York 12208, USA. [3]Howard Hughes Medical Institute, Department of Molecular, Cellular and Developmental Biology, Campus Box 347, University of Colorado at Boulder, Boulder, Colorado 80309, USA. Correspondence should be addressed to V.A. (vrambros@gmail.com) or Y.D. (yding@wadsworth.org).

RECEIVED 23 MAY; ACCEPTED 11 JULY; PUBLISHED ONLINE 17 AUGUST 2008; DOI:10.1038/NMETH.1247

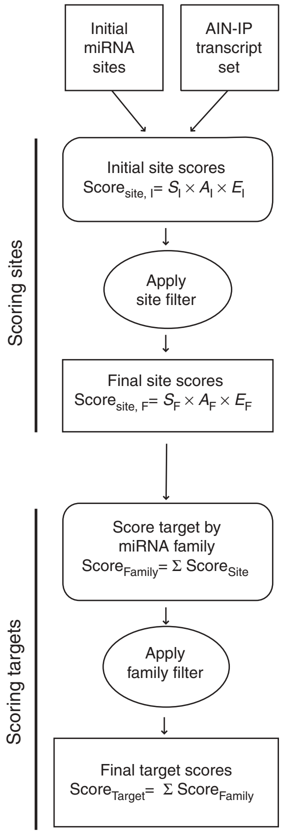

Initial miRNA sites
AIN-IP transcript set
Initial site scores
Score$_{site, I}$= $S_I \times A_I \times E_I$
Scoring sites
Apply site filter
Final site scores
Score$_{site, F}$= $S_F \times A_F \times E_F$
Score target by miRNA family
Score$_{Family}$= Σ Score$_{Site}$
Scoring targets
Apply family filter
Final target scores
Score$_{Target}$= Σ Score$_{Family}$

Figure 1 | Flowchart for the mirWIP target prediction method. An initial set of predicted miRNA binding sites was analyzed for features enriched in the 3′ UTR sequences of AIN-IP transcripts. These features were used to score individual predicted binding sites (see Methods and **Supplementary Methods**). Binding site scores were then combined into total miRNA family scores for each target, estimating the likelihood that a given transcript is regulated by a particular miRNA family. Finally, the miRNA family scores were combined into a total target score for each transcript, estimating the likelihood that a given transcript is regulated by a miRNA. S, 5′ seed matching; A, upstream structural accessibility; E, total energy; I, initial; F, final.

pairs or bulges in the seed region[10–12]. Two recent studies using immunoprecipitation of miRNA-containing ribonucleoprotein complexes indicate that only 30–45% of miRNAs associated with these complexes contain perfectly matched, conserved seed elements in their 3′ UTRs[13,14]. There is thus a need for target prediction algorithms that accurately incorporate modified 5′ seed rules.

The conservation of sequences among multiple genomes has been invaluable in identifying functional regulatory elements in genomes. Most computational methods for predicting miRNA targets include an evolutionary conservation filter, often requiring strict alignment of seed-complementary sequences across multiple genomes[7]. However, many miRNA binding sites that do not fit this strict definition could still be functionally important. For example, 40% of the verified miRNA targets in *C. elegans* reside within 3′ UTRs that align poorly between *C. elegans* and *Caenorhabditis briggsae* (for example, the *let-7* target sites in *die-1*, *lss-4* and *pha-4* (ref. 11); *let-60* (ref. 12); and *nhr-23* and *nhr-25* (ref. 15)). If the requirement for strict alignments is ignored in these cases, conserved sites for *let-7* can be found in the orthologous 3′ UTRs, indicating evolutionary selection for a functional miRNA-target interaction. Indeed, in the case of the regulatory relationship between *let-7* and *let-60* (ref. 12), the presence of *let-7* sites is conserved between worms and humans, although the sequence context of the sites is too divergent for strict alignment.

Many miRNA target prediction methods have incorporated MFE calculations to identify energetically stable base-pairing between a miRNA and its target sequence[16–21]. Some methods also include estimates of the structural accessibility of miRNA binding sites in the mRNA targets[18–21], and more recent methods join the two features into a single calculation[20,21]. Notably, the incorporation of target structure into calculations of the free energy of miRNA-target interactions can distinguish between a set of targets that tested positive for miRNA-mediated repression and a set that were refractory to miRNA-mediated repression[20]. However, current prediction methods vary widely in how energy and accessibility estimates are incorporated into their calculations. Two studies[18,19] consider accessibility of the binding sites but differ in the amount of mRNA sequence used to calculate that parameter. Two more recent studies[20,21] combine energy and accessibility calculations into a single prediction parameter but vary in the length of sequence and the method used to calculate accessibility. Further algorithm development is required to determine the optimal involvement of accessibility and binding energy in miRNA-target interactions.

Optimizing algorithms based on sequence features alone has been complicated by the lack of a large dataset of verified miRNA-target relationships. The number of targets that have been tested by rigorous genetic or reporter assays in various organisms has increased, but the assays vary in terms of how closely they model the endogenous characteristics of the interaction being tested[7]. Genome-scale datasets linking specific miRNAs to specific mRNA targets have emerged from microarray hybridization experiments that assay mRNA transcript levels after introduction of a particular miRNA by transfection[9,22]. Although these datasets have provided important insights into parameters associated with functional interactions, this approach is limited to the detection of miRNA-target interactions that result in transcript destabilization and does not identify stable, translationally repressed target mRNAs. Recently, immunoprecipitation of the RNA-induced silencing complex (RISC) has been used to identify mRNAs that stably associate with the endogenous RISC[13,14,23]. This approach provides a means of directly identifying endogenous stable complexes between miRNA-induced silencing complex (miRISC) and target mRNAs, providing large datasets of high-confidence miRNA-target interactions that can, in principle, be applied to derive target prediction algorithms of increased accuracy. One study in *C. elegans*[23] recovered 3,404 mRNA transcripts that specifically coprecipitate with the miRISC proteins AIN-1 and AIN-2. This 'AIN-IP' set of mRNA transcripts forms a biologically derived estimate for the number of genes that are targeted by miRNAs genome-wide—in this case, at least one-sixth of *C. elegans* genes.

We found several contextual features of miRNA binding sites that were enriched in sites in the AIN-IP set of transcripts: structural accessibility of target sequences, total free energy of miRNA-target hybridization and topology of base-pairing to the 5′ seed region of the miRNA. We used these features to develop a miRNA target prediction algorithm, mirWIP, that scores miRNA target sites by weighting site characteristics in proportion to their enrichment in the experimental AIN-IP set. MirWIP has improved overall

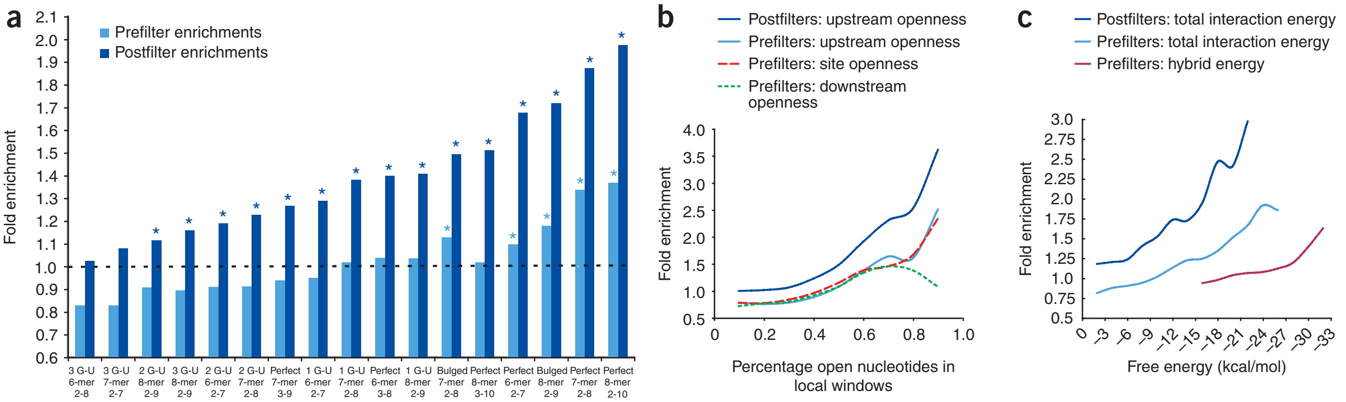


Figure 2 | Characteristics of miRNA target sites in AIN-IP transcripts. (**a**) AIN-IP transcripts are enriched for binding sites with extensive 5′ seed pairing. The horizontal axis is ordered according to final enrichment for increasing stringency in 5′ seed matches with an indicated number of G-U wobble pairs or a single bulge on the mRNA side of the duplex. The vertical axis shows the enrichment for seed matches at the indicated stringency in AIN-IP versus all other transcripts before (light blue) and after (dark blue) implementation of the site filter (see **Fig. 1** and **Supplementary Methods**). Asterisks designate significant enrichments with $P < 0.05$. (**b**) AIN-IP transcripts are enriched for binding sites that lie within structurally accessible regions. The horizontal axis shows the calculated accessibility of local sequence windows, either across the entire binding site (red dashed line), within a 25-nucleotide window upstream of the binding site (light and dark blue solid lines) or downstream (green dotted line). After applying the site filter, enrichment was calculated for upstream windows only (dark blue solid line). (**c**) AIN-IP transcripts are enriched for binding sites with favorable free energies. The conserved binding sites in AIN-IP transcripts are more likely to have favorable (negatively valued) total hybridization energies than are their counterparts in non-AIN-IP transcripts (light and dark blue lines). Also shown is hybrid energy (purple line), which reflects the stability of the final miRNA-target duplex and corresponds to the MFE. Enrichment for $\Delta G_{\mathrm{total}}$ substantially increases after applying the site filter (dark blue line).

performance compared to previous algorithms, in both recovery of the AIN-IP transcripts and correct identification of genetically verified miRNA-target relationships without a requirement for alignment of target sequences.

RESULTS

Initial target prediction

We used RNAhybrid[17] with modifications (**Supplementary Methods** online) to generate a list of all miRNA-target matches in *C. elegans* and *C. briggsae*. We filtered this initial set of raw miRNA target matches on the basis of minimal free energy, phylogenetic conservation and seed pairing configuration (**Supplementary Methods** and **Supplementary Fig. 1** online) to produce an initial list of conserved *C. elegans* miRNA binding sites (**Fig. 1**). We analyzed this set of sites to identify contextual features in AIN-IP transcripts that are enriched (**Fig. 2**) or not enriched (**Supplementary Results** and **Supplementary Fig. 2** online). We then used this information to develop an algorithm that scores miRNA binding sites and mRNA targets based on characteristics enriched in the AIN-IP set of transcripts (**Fig. 1**). We omitted the 14 experimentally verified *C. elegans* miRNA-target interactions from this analysis to retain their independence as a test of the method.

5′ seed match features enriched in AIN-IP transcripts

Extensive 5′ seed pairing shows the best enrichment for AIN-IP targets over all other transcripts assayed (**Fig. 2a**). The criterion of perfectly conserved seed matches to 8-mer blocks significantly enriches for AIN-IP targets, but these perfect 8-mers are relatively rare, residing within only 10% of the AIN-IP target transcripts. This is consistent with the occurrence of G-U wobble base-pairs and bulges in validated miRNA-target relationships and reinforces the conclusion that extensive 5′ seed pairing is neither necessary nor sufficient for reliable miRNA-target prediction. For the initial list of binding sites, it seems that perfectly matched seeds could be the only seed configurations enriched in the AIN-IP data (**Fig. 2a**). However, AIN-IP transcripts were outnumbered 3:1 by all other transcripts in this list; moreover, imperfectly paired seeds were more common than perfect matches, so these bins were more affected by the noise of false positives. With these cautions in mind, we explored the influence of other contextual features that could maximize recovery of AIN-IP transcripts, and the 14 verified interactions, while minimizing the total number of targets predicted.

AIN-IP binding sites are structurally accessible

We used the Sfold method[24] to fold whole 3′ UTR sequences plus 300 nucleotides of adjacent coding sequence for all predicted *C. elegans* transcripts. The Sfold output returns the probability that each nucleotide in the 3′ UTR is predicted to be single-stranded (that is, accessible). We used this output to calculate the average accessibility over 25-nucleotide windows around and including each potential miRNA binding site. The average structural accessibility in upstream sequence windows shows the best enrichment for AIN-IP transcripts (**Fig. 2b**).

AIN-IP binding sites are more energetically favorable

Hybridization between a miRNA and a structured mRNA target involves two major components: $\Delta G_{\mathrm{hybrid}}$, the stability (hybrid free energy) of the miRNA-target duplex, and $\Delta G_{\mathrm{disruption}}$, the cost of altering the local structure of the mRNA target[20]. For a successful hybridization, the net energy of the process, $\Delta G_{\mathrm{total}} = \Delta G_{\mathrm{hybrid}} - \Delta G_{\mathrm{disruption}}$, must be thermodynamically favorable (that is, negatively valued). The binding sites in AIN-IP structures were strongly enriched for highly favorable values of $\Delta G_{\mathrm{total}}$ (**Fig. 2c**). Because $\Delta G_{\mathrm{total}}$ is an energetic measure of the target accessibility, it is highly correlated with the average structural accessibility across the binding site, as discussed above. For this reason, the trends of enrichment were similar for these two

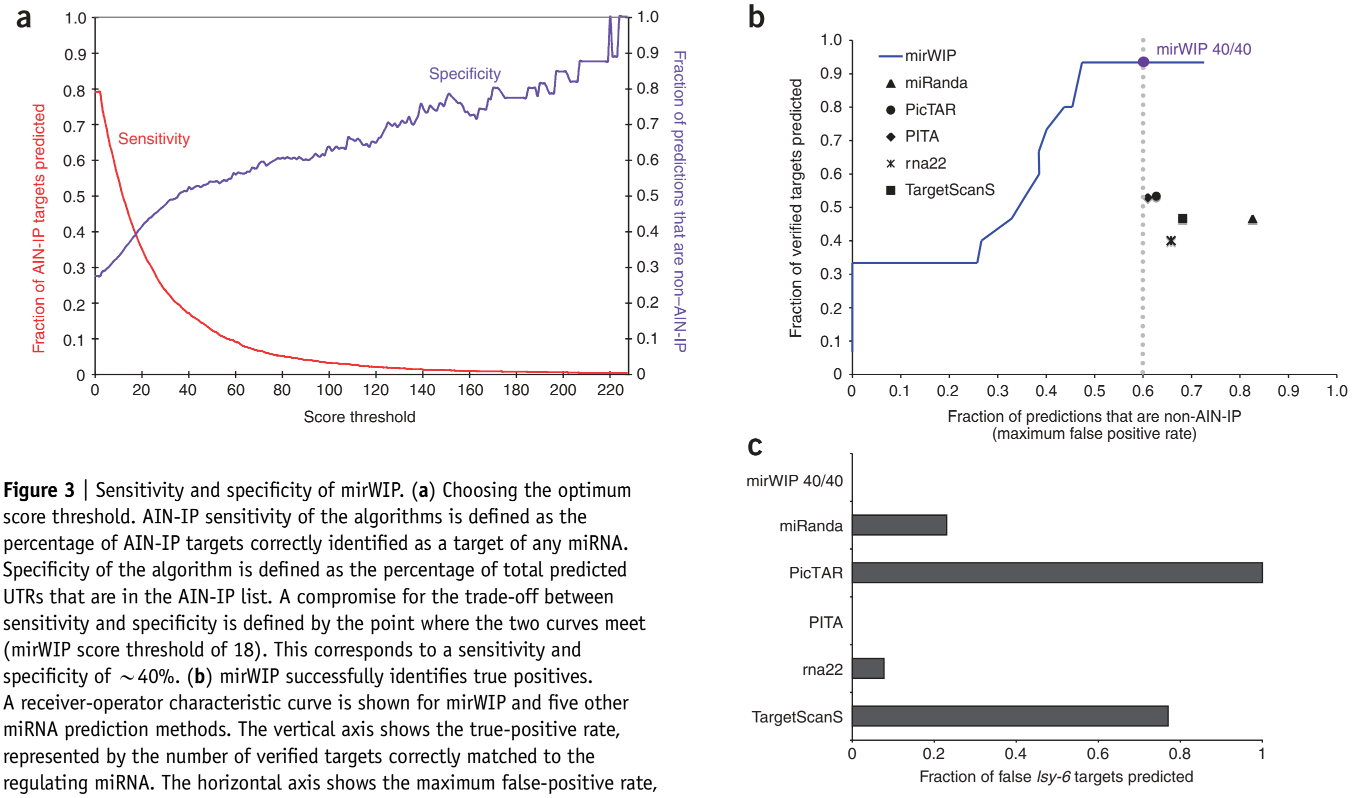


Figure 3 | Sensitivity and specificity of mirWIP. (**a**) Choosing the optimum score threshold. AIN-IP sensitivity of the algorithms is defined as the percentage of AIN-IP targets correctly identified as a target of any miRNA. Specificity of the algorithm is defined as the percentage of total predicted UTRs that are in the AIN-IP list. A compromise for the trade-off between sensitivity and specificity is defined by the point where the two curves meet (mirWIP score threshold of 18). This corresponds to a sensitivity and specificity of ~40%. (**b**) mirWIP successfully identifies true positives. A receiver-operator characteristic curve is shown for mirWIP and five other miRNA prediction methods. The vertical axis shows the true-positive rate, represented by the number of verified targets correctly matched to the regulating miRNA. The horizontal axis shows the maximum false-positive rate, the fraction of predicted UTRs that are not in the AIN-IP list. The blue line indicates the performance of mirWIP as a function of scoring threshold; the large blue dot indicates the 40% sensitivity/specificity compromise point (defined in **a**). mirWIP outperforms all five other methods by nearly doubling the true-positive rate at a lower false-positive rate (vertical gray line). (**c**) mirWIP is specific enough to reject all of the known false *lsy-6* targets (see **Supplementary Table 1**).

measurements (**Fig. 2b**,**c**). $\Delta G_{\mathrm{hybrid}}$ was substantially enriched, but to a lower degree than was $\Delta G_{\mathrm{total}}$.

miRNA target prediction using mirWIP

We used the three features that showed the best enrichment in AIN-IP targets (**Fig. 2**) to develop a miRNA target prediction scheme optimized to return AIN-IP transcripts and the verified miRNA-target relationships (listed in **Supplementary Table 1** online). This method is named mirWIP—miRNA targets by weighting AIN-IP enriched parameters. Specifically, we calculated the relative enrichments for AIN-IP targets in each of the bins for 5′ seed matching (S), upstream structural accessibility (A) and total energy (E) of the miRNA-target hybridization, $\Delta G_{\mathrm{total}}$. We used these three parameters to assign to each individual binding site three initial scoring parameters, S_{I}, A_{I} and E_{I}.

Individual binding site scores were assigned in a two-step process (**Fig. 1** and **Supplementary Methods**). After the initial scoring of all sites, we sought a mechanism to reduce noise before a second round of evaluating AIN-IP enrichment. Rather than cull all sites below a given initial score threshold, we chose to filter sites based on their overlap with higher-scoring sites in the same 3′ UTR. Accordingly, we moved a window along each UTR and retained the best nonoverlapping binding site for each position in the UTR (all overlapping binding sites were set aside). We then recalculated the relative enrichments using this filtered site dataset (shown as dark blue bars or lines in **Fig. 2**). This filtering step improved the magnitude of the relative enrichments in each bin for all three features, indicating that the filtering operation improved signal to noise. We used these postfilter weights, S_{F}, A_{F} and E_{F} (listed in **Supplementary Table 2** online), to recalculate the score for the entire set of initial miRNA sites (**Fig. 1**), including the overlapping sites previously set aside. This calculation produced the final site score, $\mathrm{Score}_{\mathrm{Site}}$ (**Fig. 1**).

After scoring all sites using the postfilter enrichments, we sought to again filter out the relatively low-scoring individual binding sites while calculating scores for 3′ UTR targets (**Fig. 1**). The optimal approach was to evaluate interactions of an entire miRNA family (as defined previously[25]) with each target (see **Supplementary Methods**). We calculated the total family score for each target by adding up all nonoverlapping site scores for each miRNA family member, separately. We then discarded any family-target interaction with a total mirWIP family score below 2.0 (see **Supplementary Methods**). Each UTR target was then given a total target score by adding up the contribution from each remaining miRNA family.

The target scores varied from 2 to 400, with the highest scores going to *lin-14* and *hbl-1*, two of the first identified miRNA targets in *C. elegans*. **Figure 3a** shows a plot of the sensitivity and specificity of our method against varying target score threshold. Sensitivity corresponds to the percentage of AIN-IP target genes successfully recovered at each threshold. Sensitivity started at 79% (instead of 100%) because some targets had no strong, conserved miRNA binding sites, as will be discussed later. Specificity represents the percentage of total predicted targets that are AIN-IP genes. For instance, with no threshold, the number of transcripts in the

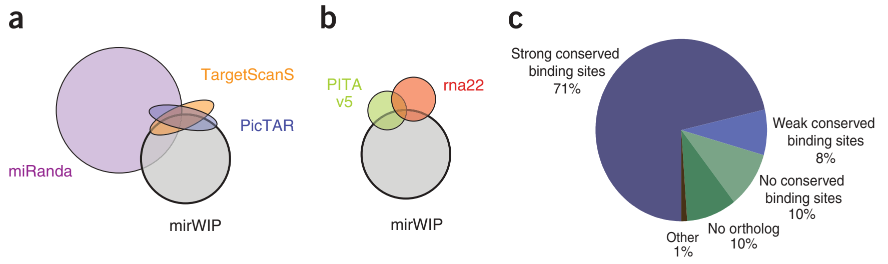


Figure 4 | Distribution of miRNA predictions. Venn diagrams showing the degree of overlap between mirWIP, miRanda, PicTAR and TargetScanS (**a**), and between mirWIP, PITA and rna22 (**b**). (**c**) AIN-IP transcripts were analyzed for site features that can account for their enrichment in the IP. Most (71%) AIN-IP transcripts contain conserved binding sites for known miRNAs. However, 29% of the AIN-IP transcripts do not have strong, conserved binding sites in their annotated 3′ UTRs.

AIN-IP list is roughly one-quarter of the total number of mRNAs examined, which corresponds to a specificity of ~27%. A compromise point can be found at a mirWIP score of 18, where the sensitivity and specificity are both ~40%. At this score level, 1,214 AIN-IP transcripts and 1,915 non-AIN-IP mRNAs were predicted as targets. This threshold easily accommodates the 14 verified *C. elegans* target genes (**Supplementary Table 1**), all of which have a score greater than 47.

mirWIP enrichments, weights and thresholds are robust

To evaluate the robustness of the optimized algorithm (in particular, to ensure that the predictions were not biased by a few high-scoring transcripts), we did a 50% cross-validation calculation. We randomly divided the data in half, derived the weights from the first half of the data, and tested how well the algorithm predicted AIN-IP versus non-AIN-IP transcripts from the remaining half of the dataset. We repeated this analysis 100 times, finding that the accuracy calculations were stable against random data shuffling, with an accuracy of 63.6% and an s.d. of 0.6%. This calculated 'accuracy' is likely to be a significant underestimate, as it was measured by the ability of the algorithm to separate AIN-IP from non-AIN-IP targets, but many non-AIN-IP targets are likely to be real.

Comparison of mirWIP performance to other methods

We compared our algorithm to the three most commonly used target prediction methods for *C. elegans*: PicTAR[16], TargetScanS[22] and miRanda[26]. We also included rna22 (ref. 6) in our comparisons as this method does not use any of the typical prediction criteria (seed matching, conservation, energy or structure). Finally, we included a recent method, PITA[21], that is similar to our technique in that PITA also uses seed, structure and energy calculations to predict target transcripts, but without sequence conservation (PITA online release 5, with the suggested threshold $\Delta\Delta G < -10$ kcal/mol). We selected these methods to show the improvements gained by using our AIN-IP–derived weights and our particular combination of contextual features.

We used two metrics to compare the performance of mirWIP to that of the other algorithms. First, we considered the ability of the algorithms to return the experimentally verified *C. elegans* miRNA-target matches listed in **Supplementary Table 1**. (Although this dataset is small, it represents the strictest test of the sensitivity of miRNA prediction methods and is a true experimental validation set for mirWIP, as these sites were not included in our enrichment analysis.) We compared this to the percentage of predicted targets that are not in either the AIN-IP or verified target list—an estimated maximum false-positive rate. A receiver-operator characteristic plot (**Fig. 3b**) shows these results, including the performance of the mirWIP algorithm at varying target score thresholds and the performance of mirWIP at the 40% sensitivity cutoff (defined in **Fig. 3a** and discussed above). The mirWIP algorithm outperformed these five prediction methods by returning more verified miRNA targets at a lower false-positive rate. The ability of mirWIP to correctly predict the weakest of the verified targets without a corresponding increase in the false-positive rate is the strongest finding of this study and highlights the utility of RISC immunoprecipitation assays in improving miRNA target prediction.

In a second estimate of algorithm specificity (**Fig. 3c**), we compared each method's recovery of a set of well-characterized false targets of *lsy-6* (ref. 27). The mirWIP algorithm does not predict any of these genes as a target of *lsy-6*, similar to PITA release 5, whereas the other four methods vary in predicting 7–100% of these interactions. This comparison may be biased against PicTAR (as compared to the other methods) as these *lsy-6* targets were specifically selected from the PicTAR predictions to show an instance where 'conserved seed' predictions fail. However, many of the validated true targets were also selected from seed-based prediction catalogs, making the true-negative comparison set as fair as the true-positive set with regard to mirWIP success rates.

Overlap among miRNA prediction methods

We next compared the overlap in predicted miRNA-target interactions for mirWIP and each of the five methods described above (**Fig. 4a,b**). We compared mirWIP to those methods that consider orthologous conservation (mirWIP, miRanda, PicTAR and TargetScanS; **Fig. 4a**) and to two methods that do not use conservation (PITA and rna22; **Fig. 4b**). miRanda predicted the largest percentage of mirWIP interactions, but it also predicted the largest number of targets overall. Notably, the overlap between mirWIP, PicTAR and TargetScanS (**Fig. 4a**) shows that mirWIP tends to include predicted targets shared by PicTAR and TargetScanS, a result of common predictions with strong seed signals. Most mirWIP predictions did not overlap with PicTAR and TargetScanS; these targets primarily show noncanonical seeds with strong structural features or functional conservation without alignment. The lack of overlap between mirWIP and rna22 is not particularly surprising, as this method differs in all aspects from the mirWIP method. However, the lack of overlap between mirWIP and PITA is notable given the similarity of these two methods.

Overall, there was only modest overlap among the six methods in the sets of miRNA-target interactions predicted. Approximately 25% of the specific miRNA-target interactions predicted by mirWIP were shared with at least one of the five other methods. However, there was better agreement among these methods in terms of the mRNAs predicted to be targeted by miRNAs in general. That is, 96% of the genes in the mirWIP catalog were also predicted to be targets of miRNAs by at least one of the other methods. In other words, these prediction methods agree about many of the genes targeted by miRNAs but disagree about which miRNA is regulating that gene. Notably, 27% of the verified miRNA-target interactions were in that set of predictions unique to mirWIP.

Analysis of falsely rejected AIN-IP targets

The mirWIP algorithm identified 79% of the AIN-IP transcripts on the basis of conserved binding sites in the 3′ UTR (**Fig. 4c**). Most of the AIN-IP transcripts that were not included by mirWIP did not pass the initial MFE and conservation filters. By relaxing the MFE filter from –15 kcal/mol to –10 kcal/mol, we found conserved binding sites for an additional 271 AIN-IP UTRs ("weak conserved binding sites" in **Fig. 4c**). Although there may be many true predictions in this group, relaxing the MFE filter would lead to a substantial increase in the false-positive prediction rate, allowing in 940 additional non-AIN-IP target UTRs and 54% of the *lsy-6* predicted sites shown to be nonfunctional[27]. The mirWIP conservation filter rejected 10% of the AIN-IP transcripts with strong binding sites for a miRNA in *C. elegans* but not in *C. briggsae*. Finally, an additional 10% of the AIN-IP genes do not have an ortholog in which to look for conserved binding sites[28]. There may be many nonconserved binding sites for known miRNAs in this group, as well as conserved binding sites for unknown miRNAs. Relaxing the already lenient orthology filter, however, would lead to an unacceptable false-positive rate as conservation is one of the strongest filters in the algorithm.

DISCUSSION

The AIN-IP set of miRISC-associated mRNA transcripts represents the largest currently available set of true miRNA targets identified from their endogenous context. This target list is not biased by selection from a particular target prediction method, allowing a fair comparison across methods. The large number of targets in the AIN-IP list allowed for a statistical analysis of both sequence and structural features associated with regulation by the miRISC complex. We found that AIN-IP transcripts are enriched for miRNA complementary sites and that certain features of the miRNA binding sites are strongly enriched. These features include a range of 5′ seed base-pairing configurations, structural accessibility of the binding site and an upstream region, and favorable total interaction energy of the miRNA-mRNA hybridization. These findings are consistent with previous reports on the importance of both canonical and noncanonical seed matches[8–11,22], target accessibility[18–21] and interaction energy[20,21].

The strongest enrichment values for structural accessibility and total hybridization energy were greater than the strongest enrichment values for seed topology. We do not believe this implies that seed matching is less predictive than the other two parameters for identifying miRNA targets, because we prescreened all potential miRNA-target binding sites to meet minimal seed criteria before we calculated enrichment values. Thus, it is possible that we are underestimating the contribution of seed matching relative to the two other parameters. We cannot predict the extent to which the enrichment scores might reflect the relative ability of each parameter to return functional miRNA binding sites. We can, however, say that the combination of these three parameters into a total scoring method outperforms a model in which one or more of these parameters is omitted or given less weight (**Supplementary Methods**).

mirWIP shows improved target prediction in *C. elegans* in several respects. First, the mirWIP method returns all 14 of the conserved, verified miRNA-target relationships without increasing the total false-positive rate beyond that of the current standard predictions. It should be emphasized that the set of 14 validated targets was not used to train the algorithm, and thus they provide an independent experimental test of the method. This list includes many noncanonical binding sites (imperfect seed matches as well as sites not conserved in aligned genomes) that cannot be identified by current target prediction methods. Second, mirWIP correctly rejects 13 targets that are predicted by other methods but have been shown to be nonfunctional *in vivo*[27]. Finally, the miRISC association of most (79%) of the AIN-IP transcripts can be explained by the existence of conserved binding sites for known miRNAs; the remaining 21% were rejected because of a lack of conserved targeting between *C. elegans* and *C. briggsae*. This scoring method can be applied to the output of any miRNA target prediction and secondary structure prediction method.

Among the mirWIP-predicted targets, 40% were identified by the AIN-IP method; 60% of the mirWIP-predicted transcripts were not stably associated with AIN proteins in the miRISC. Many of these non-AIN-IP transcripts could represent false-positive predictions by mirWIP, which would imply a lower bound of 40% for our true-positive predicted fraction. However, for several reasons, we believe that a substantial portion of these non-AIN-IP transcripts represent *bona fide* miRNA targets. First, the strict cutoff implemented in defining the AIN-IP list[23] may have removed many true targets. Second, we expect the sensitivity of the AIN-IP method to be poor for interactions that involve a small fraction of the total population of the target mRNA. For example, some interactions may occur only transiently and/or in a limited number of cells in the animal, as is the case for *lsy-6* and *cog-1* (ref. 29). Third, the AIN-IP method is likely to be most effective at recovering stable miRNA-mRNA complexes and is expected to recover unstable mRNAs much less efficiently. Some miRNAs regulate their targets on the level of mRNA stability[30], and such miRNA-mRNA complexes would be relatively short-lived and poorly detected by microarray hybridization. Finally, 4 of the 14 genetically validated miRNA targets were not in the AIN-IP list (29%). This suggests that as many as 29% of the mirWIP predictions are true miRNA targets that were not identified by AIN-IP. By this estimate, an upper bound on our positive prediction rate could be as high as 70%.

Analysis of additional experimental datasets should improve the sensitivity and specificity of mirWIP target predictions. For example, analysis of miRISC-associated RNAs from populations of developmentally staged worms or specific cell types should help reduce the noise associated with averaging regulatory interactions over all stages and tissues. Moreover, mirWIP in its current form is supported by immunoprecipitation experiments that identify transcripts by their probable association with miRNAs, but these experiments do not directly provide information about what particular miRNA or set of miRNAs is responsible for miRISC association. The immunoprecipitation of miRISC proteins from animals lacking a specific miRNA would allow us to match individual miRNAs to the targets they regulate. One such experiment[13] was conducted with a tagged version of Argonaute in *Drosophila*, significantly enriching for a small number of targets for *dme-miR-1*. Similar experiments can be applied to *C. elegans*, where a comprehensive set of miRNA mutants is available. Finally, because the miRISC immunoprecipitation approach may be biased toward the identification of stable miRNA-target complexes, miRNA-induced target destabilization can be screened using complementary datasets, such as microarray assays to identify mRNA transcripts that change in response to miRNA activity.

METHODS

miRNA target identification. We used the RNAhybrid algorithm[17] to identify the raw list of possible miRNA matches in the set of orthologous 3′ UTRs of *C. elegans* and *C. briggsae*, with a few modifications (see **Supplementary Methods**). Subsequent filtering and scoring of miRNA sites, and the derivation of methods for combining site scores to produce target (3′ UTR) scores, are described in **Supplementary Methods** and shown in **Supplementary Figure 3** online.

Structural accessibility calculations. We use the Sfold method[24] to fold 3′ UTR sequences for all *C. elegans* transcripts, plus 300 nucleotides of coding sequence adjacent to the stop codon. Details of accessibility calculations and lengths of sequences examined are given in **Supplementary Methods**.

Total interaction energy calculations. The calculations for $\Delta G_{\mathrm{total}}$ were separate from the average accessibility calculations described above, but we also used the predicted accessibilities as follows. We used the predicted structures for each binding site, calculating the energy necessary to disrupt any bound nucleotides in that region ($\Delta G_{\mathrm{disruption}}$). We then added this disruption energy to the minimal free energy, $\Delta G_{\mathrm{hybrid}}$, to obtain the total interaction energy, $\Delta G_{\mathrm{total}}$.

Statistical analysis. We estimated the significance of the prefilter enrichments for seed, structural accessibility measures and total free energy (**Fig. 2**) using Fisher's exact two-tailed contingency table. For the postfilter enrichments, which were derived from 100 random shuffles of the data, we calculated the *P* values from the *Z* score of a normal distribution. Individual *P* values for every bin are given in **Supplementary Table 2** along with a discussion of the method chosen to calculate *P* values.

Genome-wide prediction of miRNA targets. *C. elegans* genomic miRNA target predictions generated using the mirWIP algorithm are available through a web interface (http://mirtargets.org/). The mirWIP scoring method has also been implemented into the STarMir module of the Sfold package to make predictions for any miRNA-target pair from any genome of interest (http://sfold.wadsworth.org/starmir.pl/). Source code for the RNAhybrid modifications and the scoring method is available as **Supplementary Software** online.

Additional methods. Details of the initial miRNA binding site identification, calculation and statistical analysis of enrichments, alternative methods examined for scoring sites and targets, and analysis of the robustness of the calculated accuracy of mirWIP are available in **Supplementary Methods**.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank C. Hammell and all members of the Ambros lab for useful discussions and the Computational Molecular Biology and Statistics Core at the Wadsworth Center for providing computing resources for this work. This research was supported by US National Institutes of Health grants GM34028 and GM066826 to V.A., GM068726 to Y.D. and GM47869 to M. Han as well as US National Science Foundation grant DBI-0650991 to Y.D.; and the Howard Hughes Medical Institute, of which M. Han is an investigator.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Lee, R.C., Feinbaum, R.L. & Ambros, V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).
2. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
3. Jackson, R.J. & Standart, N. How do microRNAs regulate gene expression? *Sci. STKE* **2007**, re1 (2007).
4. Vasudevan, S., Tong, Y. & Steitz, J.A. Switching from repression to activation: microRNAs can up-regulate translation. *Science* **318**, 1931–1934 (2007).
5. Kloosterman, W.P. & Plasterk, R.H. The diverse functions of microRNAs in animal development and disease. *Dev. Cell* **11**, 441–450 (2006).
6. Miranda, K.C. *et al.* A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203–1217 (2006).
7. Rajewsky, N. microRNA target predictions in animals. *Nat. Genet.* **38** Suppl, S8–S13 (2006).
8. Brennecke, J., Stark, A., Russell, R.B. & Cohen, S.M. Principles of microRNA-target recognition. *PLoS Biol.* **3**, e85 (2005).
9. Lim, L.P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
10. Vella, M.C., Reinert, K. & Slack, F.J. Architecture of a validated microRNA:target interaction. *Chem. Biol.* **11**, 1619–1623 (2004).
11. Grosshans, H. *et al.* The temporal patterning microRNA let-7 regulates several transcription factors at the larval to adult transition in *C. elegans*. *Dev. Cell* **8**, 321–330 (2005).
12. Johnson, S.M. *et al.* RAS is regulated by the let-7 microRNA family. *Cell* **120**, 635–647 (2005).
13. Easow, G., Teleman, A.A. & Cohen, S.M. Isolation of microRNA targets by miRNP immunopurification. *RNA* **13**, 1198–1204 (2007).
14. Beitzinger, M. *et al.* Identification of human microRNA targets from isolated argonaute protein complexes. *RNA Biol.* **4**, 76–84 (2007).
15. Hayes, G.D., Frand, A.R. & Ruvkun, G. The mir-84 and let-7 paralogous microRNA genes of *Caenorhabditis elegans* direct the cessation of molting via the conserved nuclear hormone receptors NHR-23 and NHR-25. *Development* **133**, 4631–4641 (2006).
16. Lall, S. *et al.* A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr. Biol.* **16**, 460–471 (2006).
17. Rehmsmeier, M., Steffen, P., Hochsmann, M. & Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA* **10**, 1507–1517 (2004).
18. Robins, H., Li, Y. & Padgett, R.W. Incorporating structure to predict microRNA targets. *Proc. Natl. Acad. Sci. USA* **102**, 4006–4009 (2005).
19. Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature* **436**, 214–220 (2005).
20. Long, D. *et al.* Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.* **14**, 287–294 (2007).
21. Kertesz, M. *et al.* The role of site accessibility in microRNA target recognition. *Nat. Genet.* **39**, 1278–1284 (2007).
22. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27**, 91–105 (2007).
23. Zhang, L. *et al.* Systematic identification of *C. elegans* miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol. Cell* **28**, 598–613 (2007).
24. Ding, Y., Chan, C.Y. & Lawrence, C.E. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA* **11**, 1157–1166 (2005).
25. 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).
26. Enright, A.J. *et al.* MicroRNA targets in *Drosophila*. *Genome Biol.* **5**, R1 (2003).
27. 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).
28. Stein, L.D. *et al.* The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biol.* **1**, E45 (2003).
29. Johnston, R.J. & Hobert, O. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**, 845–849 (2003).
30. Wu, L., Fan, J. & Belasco, J.G. MicroRNAs direct rapid deadenylation of mRNA. *Proc. Natl. Acad. Sci. USA* **103**, 4034–4039 (2006).

nature | methods

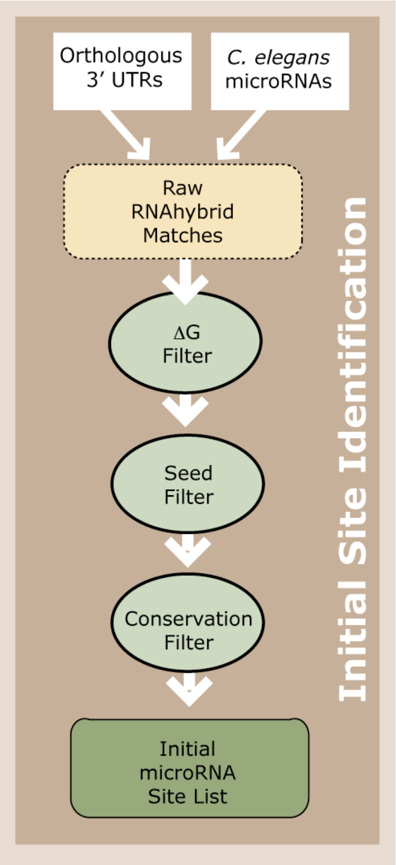
mirWIP: microRNA target prediction based on microRNA-containing ribonucleoprotein–enriched transcripts

Molly Hammell, Dang Long, Liang Zhang, Andrew Lee, C Steven Carmack, Min Han, Ye Ding

& Victor Ambros

Supplementary figures and text:

Supplementary Figure 1: Flow chart illustrating the initial microRNA binding site identification steps.



Supplementary Figure 1: Flow chart illustrating the initial microRNA binding site identification steps. As described in the Supplemental Results and Supplemental Methods sections, we began by running RNAhybrid using all C. elegans 3' UTR and microRNA sequences as input. We then implemented several liberal filters to reduce noise. These included: (1) a threshold on the hybrid interaction energy, ΔG, (2) a limit on the G:U wobble pairs and bulges allowed within the seed region and (3) a conservation filter requiring that the above two filters be passed in orthologous *C. elegans* and *C. briggsae* 3' UTRs. Details on these filters are given in the text of Supplemental Methods. The resulting list of binding sites formed the input for all of the subsequent analysis (see **Figure 1**).

Supplementary Figure 2: Contextual Features Not Correlated with Enrichment for AIN-IP Transcripts.

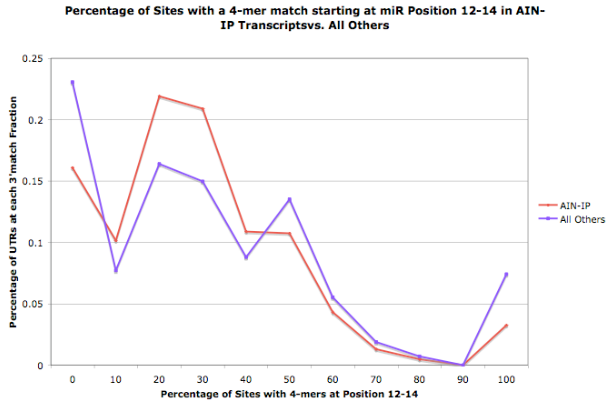


Fig 2 a

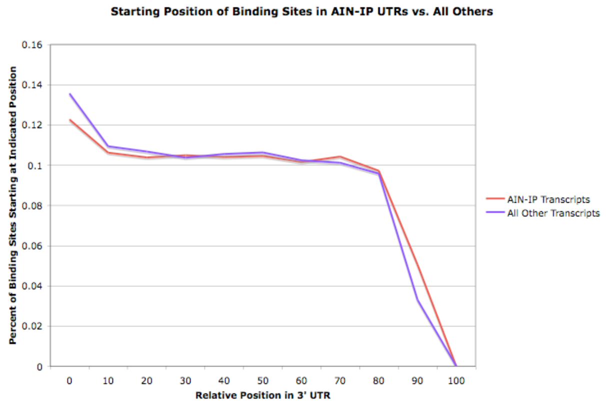


Fig 2 b

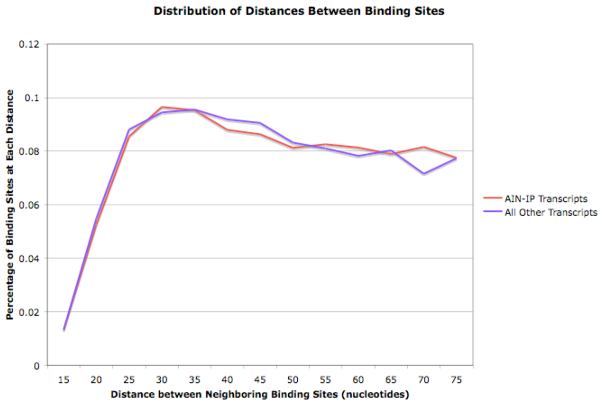
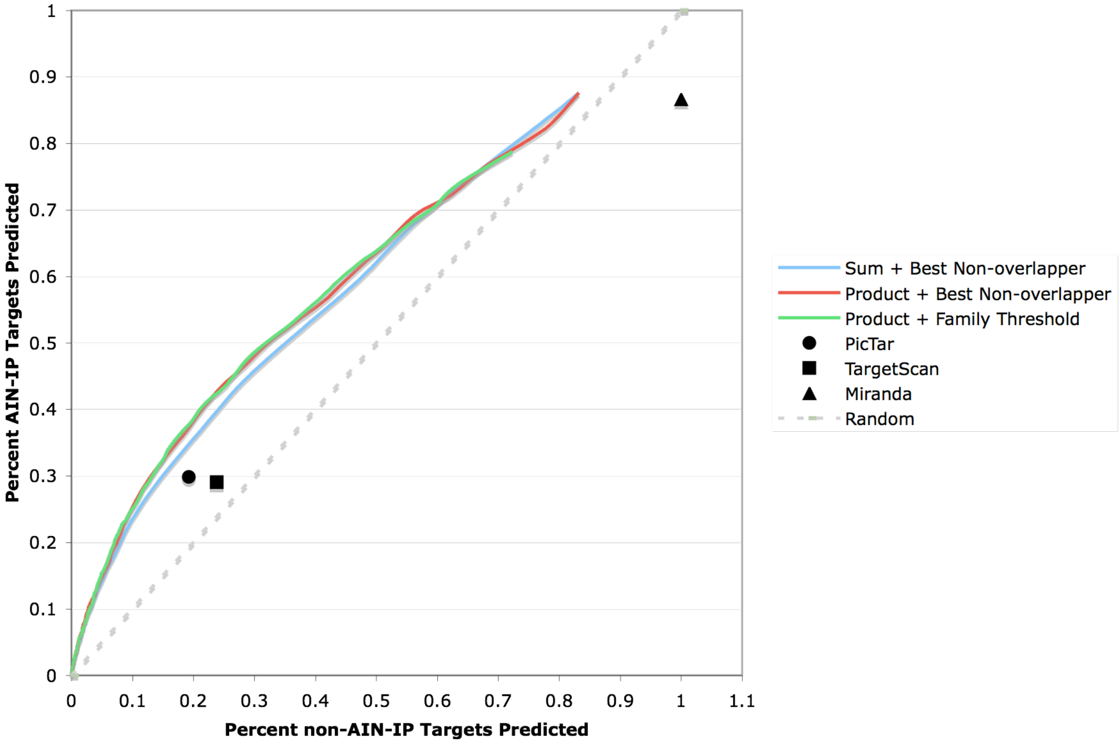


Fig 2 c

Supplementary Figure 2: Contextual Features Not Correlated with Enrichment for AIN-IP Transcripts. We evaluated several other features that have been proposed to correlate with efficacy of regulation in tissue culture transfection experiments (Grimson et al., 2007 [1]). **(a)**. A sliding window of 4 nucleotides was used to look for blocks of pairing at the 3' end of the miRNA::target duplex, across from nucleotides 12-14 of the miRNA. No significant enrichment for 3' pairing was seen for AIN-IP transcripts. **(b)**. AIN-IP transcripts show a slight bias for starting positions near the stop codon, but this bias is not significantly different from that seen in all other transcripts. **(c)**. All transcripts show a deficit in the number of binding sites that lie within about 25 nucleotides of each other, with no significant difference between AIN-IP transcripts and all others.

Supplementary Figure 3: ROC curves used to optimize mirWIP method



Supplementary Figure 3: ROC curves used to optimize mirWIP method. We optimized our algorithm using to two measures: (1) ability to return verified targets as shown in the main text and (2) ability to separate AIN-IP from non-AIN-IP targets as shown in the receiver operator characteristic (ROC) curve above. We had the option of including the enrichment factors as a linear sum of weights (blue line) or as a product of weights (red and green lines). We also had the option of choosing the best non-overlapping site or the best non-overlapping family member. For reference, we also included the performance of three highly cited methods from other groups and a random selection curve. Each mode of the mirWIP algorithm outperformed these standard methods, but the decision was made to include a product of weights with thresholds on family interactions.

Supplementary Table 1: Verified True Positive and True Negative microRNA targets in *C. elegans.*

Positive microRNA Targets	Negative microRNA Targets
lin-4::lin-14 [2]	*lsy-6*::C02B8.4
lin-4::lin-28 [3]	*lsy-6*::C27H6.3
let-7::hbl-1 [4]	*lsy-6*::C48D5.2
let-7::lin-41 [5*]	*lsy-6*::F40H3.4
let-7::let-60 [6]	*lsy-6*::F55G1.12
let-7::daf-12 [7]	*lsy-6*::F59A6.1
let-7::pha-4 [7*]	*lsy-6*::R07E3.5
let-7::lss-4 [7]	*lsy-6*::T04C9.2
let-7::die-1 [7]	*lsy-6*::T05C12.8
let-7::nhr-23 [8]	*lsy-6*::T14G12.2
let-7::nhr-25 [8]	*lsy-6*::T20G5.9
let-7::T14B1.1 [9]	*lsy-6*::T23E1.1
lsy-6::cog-1 [10*]	*lsy-6*::ZK637.13
mir-61::vav-1 [11*]	
mir-273::die-1 [12#]	

Supplementary Table 1: Verified True Positive and True Negative microRNA targets in *C. elegans.* Individual miRNA::target relationships identified in C. elegans by genetic or reporter assays. True positives are referenced in the table. True negatives were identified in a single study[13]. Asterisks (*) next to references refer to targets not identified by the AIN-IP method. [#] This microRNA is not conserved to C. briggsae, and so targets will not be found by mirWIP.

Supplementary Table 2: Lists of Enrichment Values & P-values for microRNA Binding Site Contextual Features.

Supplementary Table 2a: Seed Type Enrichments

The following tables give the absolute values for each of the pre- and post-filter enrichment values along with P-value estimates for these calculations. The data from these tables are displayed graphically in **Figure 2** and discussed in the Results section of the main text and in Supplemental Methods. P-values were calculated by Fisher's Exact two-tailed method for pre-filter enrichments and by Z-scores from a normal distribution for the post-filter enrichments. Arguments for choosing each P-value calculation method are discussed in the Statistical Analysis subsection of Supplemental Methods online. Seed configurations are listed with the number of bulges and G:U base pairs allowed in a seed of the given size at the given position (relative to the microRNA).

Seed Type	Pre-filter Enrichment	Pre-filter P-value	Post-filter Enrichment	Post-filter P-value
3 G:U 6-mer 2-7	0.830	1.99E-06	1.027	3.73E-01
3 G:U 7-mer 2-8	0.830	1.05E-08	1.081	2.68E-01
2 G:U 8-mer 2-9	0.910	1.23E-04	1.117	2.31E-02
3 G:U 8-mer 2-9	0.897	6.85E-05	1.161	6.40E-03
2 G:U 6-mer 2-7	0.913	1.41E-06	1.192	1.34E-04
2 G:U 7-mer 2-8	0.915	1.51E-04	1.230	1.16E-04
Perfect 7-mer 3-9	0.940	3.47E-01	1.268	4.61E-02
1 G:U 6-mer 2-7	0.952	4.73E-03	1.291	4.59E-12
1 G:U 7-mer 2-8	1.020	8.85E-01	1.383	1.10E-09
Perfect 6-mer 3-8	1.040	5.06E-02	1.401	7.27E-07
1 G:U 8-mer 2-9	1.038	1.81E-01	1.409	2.81E-10
Bulged 7-mer 2-8	1.130	5.74E-03	1.497	8.24E-06
Perfect 8-mer 3-10	1.020	1.00E+00	1.513	9.82E-04
Perfect 6-mer 2-7	1.099	3.28E-09	1.678	2.83E-26
Bulged 8-mer 2-9	1.180	4.48E-05	1.720	1.56E-11
Perfect 7-mer 2-8	1.340	3.98E-20	1.874	2.17E-28
Perfect 8-mer 2-10	1.370	4.70E-13	1.977	1.16E-20

Supplementary Table 2b: Structural Accessibility Enrichments

See the Supplemental Methods section for a discussion of accessibility calculations See legend for Supplementary Table 2a for more details.

Average Upstream Accessibility	Pre-filter Enrichments	Pre-filter P-values	Post-filter Enrichments	Post-filter P-values
0.1	0.783	4.93E-05	0.996	3.99E-01
0.2	0.754	2.47E-28	1.015	3.78E-01
0.3	0.782	2.79E-68	1.062	6.11E-02
0.4	0.880	3.45E-35	1.227	9.07E-13
0.5	1.073	7.24E-10	1.477	1.62E-30
0.6	1.369	1.69E-79	1.909	3.44E-32
0.7	1.633	1.07E-67	2.302	1.36E-23
0.8	1.608	2.30E-21	2.526	2.19E-11
0.9	2.505	1.99E-16	3.609	8.66E-06

Supplementary Table 2c: Total Interaction Energy (ΔG_{total}) Enrichments

See the Supplemental Methods section for a discussion of ΔG_{total} calculations. See legend for Supplementary Table 2a for more details.

Total Interaction Energy (kcal/mol)	Pre-filter Enrichments	Pre-filter P-value	Post-filter Enrichment	Post-filter P-value
-28	1.08	1.00E+00	4.00	3.81E-02
-26	1.85	1.13E-01	2.63	1.62E-01
-24	1.91	1.87E-03	2.21	2.18E-01
-22	1.66	1.13E-05	2.97	1.39E-02
-20	1.51	3.72E-09	2.41	4.44E-05
-18	1.34	7.38E-13	2.46	1.37E-08
-16	1.24	9.94E-16	1.95	4.59E-17
-14	1.22	8.96E-24	1.72	6.53E-22
-12	1.13	4.34E-13	1.73	1.87E-29
-10	1.02	2.90E-01	1.52	1.10E-20
-8	0.94	1.11E-04	1.40	8.59E-18
-6	0.90	1.37E-10	1.23	2.74E-09
-4	0.88	1.16E-11	1.20	3.78E-06
-2	0.81	5.45E-17	1.18	5.05E-04
0	0.74	1.01E-27	1.11	9.69E-02

Supplemental Results

The supplemental material included below contains data and analysis sections that we could not fit in the main text, but that we consider important for understanding: how the mirWIP analysis was done and how to repeat it on a new dataset. To that end, we include details for the initial steps taken to assemble the list of binding sites used for all analysis in the main test. We include details of enrichment analysis done for other features previously proposed to correlate with validated sites, but for which we found no enrichment in the AIN-IP dataset. We include a discussion of the remaining AIN-IP targets for which we could find no binding sites and a discussion of computational improvements to the mirWIP algorithm. Lastly, we include here most of the details for the Methods section.

Initial Target Prediction

The enriched features found by analyzing the AIN-IP dataset were used to optimize the microRNA binding sites returned by the RNAhybrid algorithm[14]. We chose RNAhybrid because this method allows us to pick potential microRNA binding sites without relying on seed-based criteria or cross-species conservation within pre-computed alignments. The miRanda algorithm[15] could have been used as our starting point, but would not have enabled us to find all microRNA:mRNA duplexes, as shown in **Figure 3b**. Several thresholds and filters were implemented through RNAhybrid and afterwards as a first step in removing noise. As shown in Supplementary **Figure 1** and described in Supplemental Methods, these included liberal filters for (1) the hybrid interaction energy, ΔG_{hybrid}, the (2) binding site seed configuration, and (3) conservation in orthologous 3' UTRs. Our initial set of all conserved microRNA:target matches in *C. elegans* contained 102,249 duplexes in a total of 8,889 UTRs. This list includes 37,226 binding sites in 2,430 AIN-IP enriched UTRs and 65,023 sites in 6,459 non-AIN-IP UTRs. Although the AIN-IP list is clearly enriched for conserved binding sites, with approximately 15 sites per UTR for the AIN-IP transcripts and 10 sites per UTR in the non-AIN-IP set, this set of microRNA:target matches is very noisy, with non-AIN-IP mRNAs outnumbering the AIN-IP mRNAs by about 3:1. Therefore an immediate goal of this study was to increase the specificity of the prediction method. The initial list of predicted UTR targets is also not completely sensitive, in that although 2,430 transcripts appear in the AIN-IP list, there remained 974 AIN-associated mRNAs for which a conserved, canonical binding site cannot be found for any known microRNA. Thus, a second goal was to determine the reasons for false negative predictions.

Contextual Features Not Strongly Correlated with AIN-IP Enrichment

We examined several other contextual features previously proposed to correlate with conserved versus non-conserved miRNA binding sites, or with efficacy of regulating transfected reporters in tissue culture. These included blocks of paired 4-mers opposite the 3' end of the miRNA[1], starting position of the binding sites with respect to the stop codon[1], and position of the binding sites relative to each other[1,16]. We found no significant enrichment for AIN-IP binding sites with any of these contextual features, as discussed below (Supplementary **Fig. 2**). Although this suggests that these features may not be generally useful for microRNA target prediction, we acknowledge that there may be classes of targets that are not well resolved by our AIN-IP vs. non-AIN-IP categories in which these additional features are enriched.

First we examined the possibility of a 3' seed region. Recently, evidence has been shown[1] for 4-mer "seeds" in the 3' end of the duplex, starting at miRNA position 12-14. We looked at the number of transcripts showing a preference for sites with 3' seed matches, seeing no significant difference between sites in AIN-IP transcripts and all other genomic UTR predictions (Supplementary **Fig. 2a**). This lack of a strong signal for 3' matching is likely due to the fact that contributions from the 3' end of the miRNA have already been taken into account by folding the entire miRNA::target duplex before declaring a match, whereas the Grimson et al.[1] study looked only for conserved 5' seed matches, then analyzed that group for additional 3' pairing. In other words, pairing the 3' end of the miRNA is important for stability of the miRNA::target duplex, which is best analyzed by looking at the duplex as a whole rather than searching for positioned 4-mers.

Next, we analyzed the positions of conserved binding sites both relative to the stop codon and relative to each other. We find a small bias toward binding sites near the stop codon in 3' UTRs (Supplementary **Fig. 2b**) but this bias is not significantly different for AIN-IP transcripts relative to all other transcripts in the genome. The lack of a difference between AIN-IP transcripts and all others suggests that there may be a general gradient of coding bias across the length of the transcripts, with conserved binding sites preferentially occurring near the coding region. This feature may be affected by the relatively short length of 3' UTRs in *C. elegans* compared to other model organisms. We found a deficit in the number of binding sites that occur too closely together (less than 25 nucleotides apart, Supplementary **Fig. 2c**) and a slight decrease in the number of sites spaced greater than 40-45 nucleotides away from each other. A previous report[16] found that an optimal distance of 13-35 nucleotides showed correlation with efficacy of reporter down-regulation. However, our dataset showed no significant difference in the distribution of relative miRNA distances between the AIN-IP dataset and all other transcripts.

AIN-IP transcripts with no predicted microRNA binding sites

The Results section of the main text contains an analysis of AIN-IP transcripts for which no conserved, canonical microRNA binding sites can be found by the mirWIP method. We found that a majority of these transcripts contain non-conserved microRNA binding sites. However, there are still 37 transcripts in the AIN-IP list that cannot be explained by conserved or non-conserved binding sites for any known microRNAs. While some of these transcripts could be targeted by microRNAs that have not yet been identified in C. elegans, examination of the distribution of UTR lengths in this group suggests that most of these remaining UTRs have uncommonly short 3' UTRs, typically 10-100 nucleotides. Considering the need for other 3' regulatory elements, such as polyadenylation signals, this leaves little space for microRNA regulatory elements. Poor annotation of 3' UTRs could be one reason for the false negative predictions of these transcripts. The ModEncode consortium has undertaken a project to map all of the 3' UTRs in *C. elegans* (http://utrome.org), reflecting the fact that UTR endpoints remain in need of better curation. Better UTR annotation might result in a significant shift in the number and length of annotated UTRs in *C. elegans*, improving microRNA target predictions. In conclusion, lack of conservation appears to be the main reason for false rejection of binding sites in AIN-IP UTRs. The group of AIN-IP transcripts for which no strong microRNA binding site can be found is a small percentage of the total.

Computational Improvements

Other potential computational refinements to mirWIP could include exploration of additional modes of combining enrichment scores than the two tested here (additive and multiplicative). We found that the multiplicative method performed better than an additive method of incorporating enrichment weights. The multiplicative scoring method for mirWIP implies that the variables (i.e., the three enriched features) are interactive and may not act independently of each other. A linear scoring method implicitly assumes that the variables are independent of each other and that the effects are additive. We chose not to undertake a more extensive evaluation of the interdependence of the variables at this time because the dataset is not quantitative enough to statistically explore and characterize the potential dependency between the variables. Such an analysis would require knowledge of the contribution of each microRNA binding site to the total regulation of the target mRNA. We chose to simply evaluate the ability of these two major classes of scoring models to optimally select targets.

Supplemental Methods

RNAhybrid

RNAhybrid[14] works by recursively returning the most energetically favorable duplex anywhere on a UTR for a query microRNA, then blocking out that entire region of the match on the UTR. The program continues until no more matches can be found that pass a given minimal free energy (MFE) threshold. Occasionally, this recursive search and “blackout” cycle resulted in selecting a poor seed site over an overlapping perfect seed match with a slightly poorer MFE. We modified the RNAhybrid source code to turn off this blackout mechanism then filtered the results to remove any redundant predictions.

While the RNAhybrid algorithm does contain parameter settings designed to allow G:U base pairs in the seed region, it does not contain parameter settings designed to allow a specified number of bulges in the seed region. Therefore, we ran RNAhybrid in its most liberal form (without a forced seed helix), to produce a list of raw matches that was then post-processed as described below through three filtering steps (hybridization energy, seed topology and conservation) to yield the initial microRNA site list (Supplementary **Fig. 1**).

Hybridization energy filter

First, we implemented a hybrid duplex energy filter. To capture all verified targets, we set a limit for the hybrid minimal free energy (MFE) of the duplex interaction at $\Delta G_{hybrid} \leq -15$ kcal/mol (just under the weakest energy needed to capture all verified targets). This filter was implemented directly in the parameter set used for the RNAhybrid scripts.

Seed topology filter

A generous seed topology filter was designed to remove from the set of raw matches any predicted duplexes with very poor base pairing to the 5' seed region of the microRNA, but to nevertheless capture the experimentally verified interactions that rely on non-canonical seed pairing. Specifically, to capture the relationship between *let-7* and *let-60* [6], we allowed up to 3 G:U wobble pairs in a 6-mer beginning at microRNA position 2. To capture both of the verified *let-7:lin-41* sites [17], we allowed up to 1 bulge on the mRNA side of the duplex in a 7-mer beginning at microRNA position 2. To capture the imperfectly paired *lin-4:lin-14* sites shown to be required for proper *lin-14* regulation [18], we allowed perfect 6-, 7-, and 8-mers beginning at microRNA position 3. In the *lin-14* study, a microRNA:target configuration was proposed that included a single bulged nucleotide on the microRNA side of the duplex in the 5' seed region. However, the *lin-4* microRNA sequence has since been revised [19], removing 2 nucleotides from the 5' end of the mature microRNA. With this change, the most energetically favorable structure shows a perfect 6-mer beginning at microRNA position 3, rather than a bulged seed. There are 12 other verified targets, listed in **Table 1** of the main text. In the case of *nhr-23* and *nhr-25*, the authors identified only *nhr-25* as a target of *let-7* and *mir-84* [8]. While only *nhr-25* contains perfect seed matches to the *let-7* family in its 3'UTR, *nhr-23* also contains strong non-canonical seed matches to *let-7* and is equally supported, genetically. In our analysis, an N-mer match to the seed region always begins with microRNA position 2 or 3. We chose not to begin the match with microRNA nucleotide 1, or adenosine anchored matches as in TargetScanS [1] as this did not enrich for AIN-IP targets beyond those that already matched positions 2-10 (data not shown).

Conservation filter

Finally, we implemented a relaxed orthology filter. We required an orthologous match in both *C. elegans* and *C. briggsae,* but we did not require that the conserved duplexes lay within aligned blocks, nor that the structure of the duplexes matched in both *elegans* and *briggsae*. Any match for a single microRNA that passed the above filters in both orthologous UTRs was counted as a conserved site. A similarly relaxed definition of conservation without alignment has previously been used for regulatory element prediction[20].

Structural Accessibility Calculations

Accessibility predictions from a single MFE structure may not be ideal for an mRNA, which may exist in any one of multiple structural configurations with a similar MFE. To address this issue, we use the Sfold method[21] to fold 3' UTR sequences for all *C. elegans* transcripts, plus 300 nucleotides of coding sequence adjacent to the stop codon. The additional coding sequence was added to avoid biasing our calculations toward open regions at the beginning of the 3' UTR and to include interactions between 3' UTR nucleotides and the adjacent coding sequence. The Sfold output returns the probability of remaining unpaired, for each nucleotide in the 3' UTR, averaged across 1000 structures with energies similar to that of the MFE structure, as discussed in[21]. We used those probabilities to then calculate the average accessibility of 3 regions: (1) a 25 nucleotide window upstream of the binding site, (2) across the entire length of the binding site, and (3) a 25 nucleotide window downstream of the binding site. We explored alternate window sizes from 10-50 nucleotides, finding the best signal-to-noise ratio at 25 nucleotides (data not shown).

Total Interaction Energy Calculations

The calculations for total interaction energy, ΔG_{total}, are separate from the average accessibility calculations performed above but do also use the predicted accessibilities, as follows. Specifically, 1000 structures are sampled from the partition function for all possible structures of a given mRNA sequence (folded using the entire 3' UTR plus 300 nucleotides of coding sequence, as above). We then look at the predicted structures in the region of each binding site individually, calculating the energy necessary to disrupt any bound nucleotides in that region, $\Delta G_{disruption}$, for each of the 1000 structures. These calculated disruption energies are then averaged to obtain a single average $\Delta G_{disruption}$ energy for each binding site. This average disruption energy is then added to the energy gained by forming the microRNA:mRNA heteroduplex to obtain the total interaction energy, ΔG_{total}. While it is possible that sampling from the partition function could introduce errors, it is just as possible that using the entire partition function does not accurately model the range of structures that are most commonly adopted by a given mRNA. This would also introduce errors by calculating disruption energies for structures that do not exist, *in vivo*.

Identification of Optimal Weighting

Scoring enrichment of site features

The formula used for relative enrichments throughout this study is as follows:

$$\text{Relative Enrichment}_{\text{bin = i}} = (n_{\text{IP,i}} / N_{\text{IP}}) / (m_{\text{non-IP,i}} / M_{\text{non-IP}})$$

Where: $n_{\text{IP,i}}$ is the number of objects from the AIN-IP list in the i^{th} bin and; N_{IP} is the total number of objects in the AIN-IP list; $m_{\text{non-IP,i}}$ is the number of objects in the non-AIN-IP list in the i^{th} bin; and $M_{\text{non-IP}}$ is the total number of objects in the non-AIN-IP list. In other words, relative enrichments for each bin measure how enriched that bin is relative to the total number of objects in each list, separately. All enrichment values, as well as P-values for each bin are given in Supplementary **Table 2**.

Site S/N filter for improving enrichments

To obtain the optimal seed, energy and accessibility weights for scoring sites, we undertook a second round of weighting based on a subset of the data that had been selected for a higher signal/noise (S/N). Specifically, we first filtered the raw dataset by eliminated any overlapping binding sites from consideration. After using the initial (raw) enrichment factors as weights to score all sites by their individual seed, structure and energy configurations, we then isolated the best of these sites by selecting only the best non-overlapping binding sites from the UTR (discarding any overlapping sites with a lower score than the top site for that position on the UTR). This produced a working set of "best non-overlapping sites" for all microRNAs. We then recalculated the enrichment factors (*S*, *E* and *A*) using just the set of best non-overlapping sites. These enrichment values after imposing the site S/N filter are shown as dark blue lines in **Figure 2**. These post-filter enrichment values were employed in a subsequent rescoring of the entire initial microRNA site list (**Fig. 1**).

This non-overlapping site S/N filter performed two functions. First, any bias from large families of microRNAs would be removed by eliminating the presence of redundant counting for each microRNA family member. Second, we observed that the best binding sites in AIN-IP UTRs were much better than the best binding sites in non-AIN-IP UTRs (by the 3 metrics of seed, structure, and energy, described above). That is, by eliminating all overlapping binding sites, we exaggerated the difference between AIN-IP and non-AIN-IP targets for all enrichment bins. This is likely due to the fact that the set of best non-overlapping binding sites in AIN-IP UTRs is enriched for true targets, while the set of best non-overlapping sites in non-AIN-IP UTRs is still dominated by noise.

Calculating final mean enrichment values

To evaluate the robustness of the calculated enrichment factors after implementation of a S/N filter, we re-calculated the enrichment factors for these "best non-overlapping sites" for 100 iterations of a random 50% data selection, obtaining mean enrichment factors, and errors bars based on the standard deviation from the mean for all three features.

As seen in Supplementary **Table 2**, nearly every feature is enriched in the post-filter calculations because the total number of sites per UTR in the "best AIN-IP" list has many more miRNA binding sites than the "best non-AIN-IP" list does, even though the AIN-IP list contains far fewer genes overall. This reflects a higher density of binding sites in the AIN-IP UTRs. Lastly, the fold enrichments are relative enrichments: percent enrichment of AIN-IP sites in a bin out of all AIN-IP binding sites relative to the percent enrichment of all non-AIN-IP sites in that bin out of all non-AIN-IP binding sites.

Statistical analysis of enrichment

The significance of the pre-filter enrichments for seed, structural accessibility measures, and total free energy, listed in Supplementary **Table 2**, can be assessed using Fisher's Exact two-tailed contingency table. This calculates the probability that a random division of objects into the given categories would show the same or greater relative enrichment/depletion. Fisher's test does not require that the two variables be normally distributed. Since we have no prior knowledge about whether each category would be enriched or depleted, we chose the two-tailed P-value, which is more conservative.

For the enrichments in seed configuration obtained prior to implementation of the site S/N filter, the null hypothesis for AIN-IP category enrichment is rejected at the $P < 0.05$ significance level for the following categories: perfect 6-, 7-, and 8-mers beginning at microRNA position 2, and bulged 7- or 8-mers. While the relative depletion of poor seeds in the AIN-IP relative to non-AIN-IP list is statistically significant, we believe this is due to the enormous amount of noise in both AIN-IP and non-AIN-IP bins for these categories. In other words, this reflects merely the high false positive rates associated with poor seed matching, but not a lack of poorly matched seeds in functional microRNA response elements. This is illustrated by the increase in relative enrichments seen after applying the first set of site filters. For ΔG_{total} and structural availability, we also chose to use Fisher's exact test. Even though ΔG_{total} and openness are continuous variables, these were binned for the purposes of calculating relative enrichments. Therefore, each bin also meets the criteria for using Fisher's contingency table. We found that the null hypothesis was rejected for ΔG_{total} bins from -12 to -24 kcal/mol. Bins of $\Delta G_{total} < -24$ kcal/mol did not have enough counts to be statistically significant. Similarly, for structural availability, P-values were less than 0.05 for bins of 50-90% average open nucleotides in the upstream window.

For the post-filter enrichments, which were derived from 100 random shuffles of the data, we calculated the P-values from the Z-score of a normal distribution. That is, random data shuffling should produce random noise in the calculated enrichments. We checked that this was true by first looking at a small number of random shuffles (10), observing that the standard deviations were relatively small (early convergence on the mean), then repeating the entire algorithm for a total of 100 iterations. The P-value for rejection of the null hypothesis (that the feature is not enriched) is the distance of each enrichment from the null hypothesis (relative enrichment – 1), divided by the standard deviation from the mean – in other words, a Z-score.

Scoring microRNA Sites, Families and Targets

Testing alternative modes of combining weights for different parameters

Contributions from the three enriched features: seed topology (represented by the variable S), structural accessibility for the 25 nt region upstream of the site (A), and total interaction energy of the microRNA:target hybridization (E), were varied in combination until the optimal prediction of AIN-IP targets was found. We calculated the best method by evaluating its performance on a ROC curve (Supplementary **Fig. 3**) plotting sensitivity to AIN-IP targets versus sensitivity to non-AIN-IP targets (one estimated measure of false positives). Specifically, we first evaluated combining the weights as a linear sum, by adding the enrichment scores of each feature multiplied by a non-negative weight, i.e., $w_1 \times S + w_2 \times \mathrm{A} + w_3 \times E$, where $w_1 + w_2 + w_3 = 1$. We varied the weights (w_1, w_2 and w_3) in increments of 1/9, from 0.0 to 1.0. We also examined the performance of an alternative version of the algorithm, where we multiplied the enrichments of each feature together ($S \times A \times E$) to see if multiplying could exaggerate the difference between high and low scores. The optimal scheme for combing enrichment values was a product (Site score = $S \times A \times E$), as discussed below.

Testing alternative signal/noise (S/N) filtering strategies for scoring UTRs

We tested the performance of the algorithm with various S/N filters applied during the scoring of target UTRs, prior to combining enrichment values by product ($S \times A \times E$), or by addition (as described above). Performance was evaluated by the method of optimizing the area under the ROC curve (Supplementary **Fig. 3**). These S/N filters consisted of removing overlapping binding sites that scored poorly.

The first class of overlap filter that we tested retained only the highest scoring of all overlapping sites for any individual microRNA. This filter was implemented using either of two options for calculating individual site scores -- either a linear sum of feature weights (option 1, below), or a product of feature weights (options 2 and 3, below). The performance of these alternative approaches was evaluated using a ROC plot (Supplementary **Fig. 3**).

Option 1: In the first option, we include a linear sum of feature weights: Site score = $w_1 \times S + w_2 \times A + w_3 \times E$. Where S is the enrichment for seed topology (**Fig. 2a**), A is the enrichment for upstream structural accessibility (**Fig. 2b**) and E is the total interaction energy, ΔG_{total}, for microRNA hybridization (**Fig. 2c**). We then eliminated all overlapping binding sites with a lower score, and calculated a UTR score = Sum of the best non-overlapping site scores. We then varied the weights (w_1, w_2, w_3) in increments of 1/9, from 0 to 100% until an optimum area under the above ROC curve was met. An optimum area under the curve was found to be 62.9% for weights of ($w_1 = 0 \times S$) + ($w_2 = 2/9 \times A$) + ($w_3 = 7/9 \times E$). This method rejected the verified *let-7:pha-4* interaction.

Option 2: In the second option, we considered each site score as a product of feature weights; Site score = $S \times A \times E$. We then eliminated overlapping binding sites with a lower score, and calculated UTR score = Sum of the best non-overlapping site scores. This involved no optimization of feature weighting. The area under the curve came out to 63.9% (red line, Supplementary **Fig. 3**), which is better but essentially similar to the first option. This method also rejected the verified *let-7:pha-4* interaction.

Option 3: Lastly, microRNAs were scored as a product of feature weights (as in Option 2), but are then grouped by seed family (as defined in Ruby et al. 2007[19]) before screening out overlapping sites. We counted only the highest scoring site for each overlapping family member, allowing for the possibility of overlapping sites between non-family members. Allowing overlap of sites between non-family members still produced a significant amount of noise (many non-AIN-IP sites). Therefore, we included an additional step to further reduce low-scoring binding sites. We screened out low-scoring microRNA families on each target by implementing a score threshold for each microRNA family before determining the final UTR score. This threshold was varied from 0 to 10 to find the optimum. We found that a family interaction threshold of "2" performed best with an effective area of the ROC curve of 63.8% (green line, Supplementary **Fig. 3**). This ROC accuracy measure was very similar to the two previous methods discussed above. Importantly, this set of options did not reject any of the verified targets, including *let-7:pha-4.* Therefore, Option 3 was the method selected for the "high stringency" target prediction by mirWIP.

Enrsuring the mirWIP method is Robust

Lastly, we examined whether the mirWIP method was robust to sub-selections of the AIN-IP dataset. This ensures that our optimization procedures reflect trends inherent in the entire dataset and are not dependent on a small number of high scoring targets. To that end, we randomly divided the data in half, derived the enrichment weights from binding sites in the first half of the dataset, applied those weights to the second half of the data, then calculated the area under the curve from a ROC plot showing only the second half of the data. We repeated this 100 times, finding a mean effective accuracy of 63.6% with a standard deviation of 0.6%. The accuracy calculations were robust, meaning they reflect a general trend of the entire dataset.

Supplemental References

[1] Grimson, A. et al., MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27** (1), 91 (2007).

[2] Lee, R. C., Feinbaum, R. L., and Ambros, V., The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75** (5), 843 (1993).

[3] Moss, E. G., Lee, R. C., and Ambros, V., The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. *Cell* **88** (5), 637 (1997).

[4] Abrahante, J. E. et al., The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. *Dev Cell* **4** (5), 625 (2003).

[5] Reinhart, B. J. et al., The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403** (6772), 901 (2000).

[6] Johnson, S. M. et al., RAS is regulated by the let-7 microRNA family. *Cell* **120** (5), 635 (2005).

[7] Grosshans, H. et al., The temporal patterning microRNA let-7 regulates several transcription factors at the larval to adult transition in C. elegans. *Dev Cell* **8** (3), 321 (2005).

[8] Hayes, G. D., Frand, A. R., and Ruvkun, G., The mir-84 and let-7 paralogous microRNA genes of Caenorhabditis elegans direct the cessation of molting via the conserved nuclear hormone receptors NHR-23 and NHR-25. *Development* **133** (23), 4631 (2006).

[9] Lall, S. et al., A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol* **16** (5), 460 (2006).

[10] Johnston, R. J. and Hobert, O., A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. *Nature* **426** (6968), 845 (2003).

[11] Yoo, A. S. and Greenwald, I., LIN-12/Notch activation leads to microRNA-mediated down-regulation of Vav in C. elegans. *Science* **310** (5752), 1330 (2005).

[12] Chang, S. et al., MicroRNAs act sequentially and asymmetrically to control chemosensory laterality in the nematode. *Nature* **430** (7001), 785 (2004).

[13] Didiano, D. and Hobert, O., Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat Struct Mol Biol* **13** (9), 849 (2006).

[14] Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R., Fast and effective prediction of microRNA/target duplexes. *Rna* **10** (10), 1507 (2004).

[15] Enright, A. J. et al., MicroRNA targets in Drosophila. *Genome Biol* **5** (1), R1 (2003).

[16] Saetrom, P. et al., Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Res* **35** (7), 2333 (2007).

[17] Vella, M. C., Reinert, K., and Slack, F. J., Architecture of a validated microRNA::target interaction. *Chem Biol* **11** (12), 1619 (2004).

[18] Ha, I., Wightman, B., and Ruvkun, G., A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. *Genes Dev* **10** (23), 3041 (1996).

[19] Ruby, J. G. et al., Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127** (6), 1193 (2006).

[20] Elemento, O. and Tavazoie, S., Fastcompare: a nonalignment approach for genome-scale discovery of DNA and mRNA regulatory elements using network-level conservation. *Methods Mol Biol* **395**, 349 (2007).

[21] Ding, Y., Chan, C. Y., and Lawrence, C. E., RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *Rna* **11** (8), 1157 (2005).