Cell

A Pattern-Based Method for the Identification of MicroRNA Binding Sites and Their Corresponding Heteroduplexes

Kevin C. Miranda,[1,4] Tien Huynh,[1,4] Yvonne Tay,[2,4] Yen-Sin Ang,[2,4] Wai-Leong Tam,[2] Andrew M. Thomson,[2] Bing Lim,[2,3,5] and Isidore Rigoutsos[1,5,*]

[1] Bioinformatics and Pattern Discovery Group, IBM Thomas J. Watson Research Center, Yorktown Heights, P.O. Box 218, NY 10598, USA

[2] Stem Cell and Developmental Biology, Genome Institute of Singapore, 60 Biopolis Street, Genome #02-01, 138672, Singapore

[3] Beth Israel Deaconess Medical Center, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02215, USA

[4] These authors contributed equally to this work.

[5] These authors contributed equally to this work.

*Contact: rigoutso@us.ibm.com

DOI 10.1016/j.cell.2006.07.031

SUMMARY

We present *rna22*, a method for identifying microRNA binding sites and their corresponding heteroduplexes. *Rna22* does not rely upon cross-species conservation, is resilient to noise, and, unlike previous methods, it first finds putative microRNA binding sites in the sequence of interest, then identifies the targeting microRNA. Computationally, we show that *rna22* identifies most of the currently known heteroduplexes. Experimentally, with luciferase assays, we demonstrate average repressions of 30% or more for 168 of 226 tested targets. The analysis suggests that some microRNAs may have as many as a few thousand targets, and that between 74% and 92% of the gene transcripts in four model genomes are likely under microRNA control through their untranslated and amino acid coding regions. We also extended the method's key idea to a low-error microRNA-precursor-discovery scheme; our studies suggest that the number of microRNA precursors in mammalian genomes likely ranges in the tens of thousands.

INTRODUCTION

MicroRNAs are short RNAs that either degrade or disrupt translation of mRNA transcripts post-transcriptionally in a sequence-specific manner (Filipowicz, 2005; Hammond, 2005; Hannon, 2002; Mattick and Makunin, 2005). Early studies suggested only a limited role for RNAi but the subsequent discovery of many endogenously encoded microRNAs (Lagos-Quintana et al., 2001, 2003; Lee and Ambros, 2001) pointed toward the possibility of this being a more general control mechanism (Hobert, 2004). More recent evidence has added support to the hypothesis that a much wider spectrum of biological processes may be affected by RNAi (Croce and Calin, 2005).

The computational methods for detecting microRNA binding sites that have been published thus far are diverse. One group of approaches (Enright et al., 2003; John et al., 2004; Kiriakidou et al., 2004) is based on variants of the dynamic programming solution to the "local suffix alignment" problem (Gusfield, 1997). A second group is "signature-based" with the signature derived from the first few nucleotides (typically positions 2 through 8 inclusive) of the microRNA's 5′ region; this signature is referred to as the "seed" or "nucleus" (Krek et al., 2005; Lewis et al., 2003, 2005; Rajewsky and Socci, 2004). Other schemes use hidden Markov models to find seed matches (Stark et al., 2003) or calculate binding interactions for every offset of the target sequence of the microRNA and subselect those relative placements that are deemed significant according to an energy-based statistical measure (Rehmsmeier et al., 2004). All these computational methods can be applied to individual genomes in isolation; however, enforcing the conservation of a potential binding site at orthologous positions across multiple species has been used as a typical filtering criterion prior to reporting results. The inherent difficulty of the microRNA target prediction problem is underscored by the fact that the number of confirmed heteroduplexes remains small by comparison to the expended effort and by the observation that predictions made by the various algorithms generally have little overlap (Rajewsky, 2006).

In this study, we present *rna22*, a pattern-based approach for the discovery of microRNA binding sites and their corresponding microRNA/mRNA complexes. *Rna22* has high sensitivity, is resilient to noise, and can be applied to the analysis of any genome without requiring genome-specific retraining. It is also distinct from

previously reported methods in that it obviates the use of a cross-species sequence conservation filter, thus allowing the discovery of microRNA binding sites that may not be present in closely related species. *Rna22* can identify putative microRNA binding sites without a need to know the identity of the targeting microRNA; this permits the identification of binding sites even if the targeting microRNA is not among those currently known. Using luciferase reporter assays, we offer new insights regarding the principles that govern the recognition of targets by microRNAs, provide experimental support that for a large fraction of *rna22*-predicted targets the observed repression will be substantial, and present evidence that some microRNAs may have as many as a few thousand targets. We also describe a methodological extension that allows the prediction of microRNA precursors from genomic sequence and use it to estimate the number of microRNA precursors in several genomes. The results of this work suggest that in a given genome the true numbers of microRNA precursors, microRNA binding sites and affected gene transcripts may be substantially higher than currently hypothesized and that, in addition to 3′UTRs, numerous binding sites likely exist in 5′UTRs and CDSs.

RESULTS

A flowchart showing the various steps of the *rna22* method can be seen in Figure 1. Below, we present and discuss each of the steps in detail.

Input Preparation

We processed the 644 mature microRNA sequences contained in Release 3.0 (January, 2004) of RFAM (Griffiths-Jones et al., 2003). We worked with a 2-year-old release of the RFAM database in order to gauge the ability of our method to extrapolate from a small repository of available knowledge. Prior to processing, we removed identical and near-duplicate entries from this collection using a previously described scheme (Rigoutsos et al., 2006) that is based on BLASTN (Altschul et al., 1990): no two remaining sequences from the final set of 354 agree on more than 90% of their positions.

Pattern Discovery

The *Teiresias* algorithm (Rigoutsos and Floratos, 1998) was used to discover variable-length motifs (“patterns”) in the mature microRNA sequences of the cleaned-up input. These motifs comprise a *minimum* of $L = 4$ nucleotides, have at least 30% of their positions specified (i.e., $W = 12$) and appear *a minimum* of $K = 2$ times in the processed input (see the Supplemental Data available with this article online, regarding the values of L, W, K). The algorithm guarantees the reporting of *all* composition-maximal and length-maximal patterns that satisfy the given parameters.

Conserved sequence segments of the type discussed here are typically represented by regular expressions with varying degrees of descriptive power (Brazma et al., 1998). In this analysis, we use expressions that comprise a combination of *literals* (solid characters from the alphabet of permitted symbols), *wildcards* (each denoted by “.” and representing *any* character), and, *sets of equivalent literals* with each set being a small number of symbols any one of which can occupy the corresponding position. One such pattern is [AT][CG].TTTTT[CG]G..[AT], all instances of which have their first position occupied by either an A or T, their second position by a C or G, their third position by any nucleotide, their fourth position by a T, etc. The distance between two consecutive occupied positions is unchanged across all instances of the pattern (= “rigid patterns”) (see Supplemental Data for a discussion). Using actual genomic data, we trained

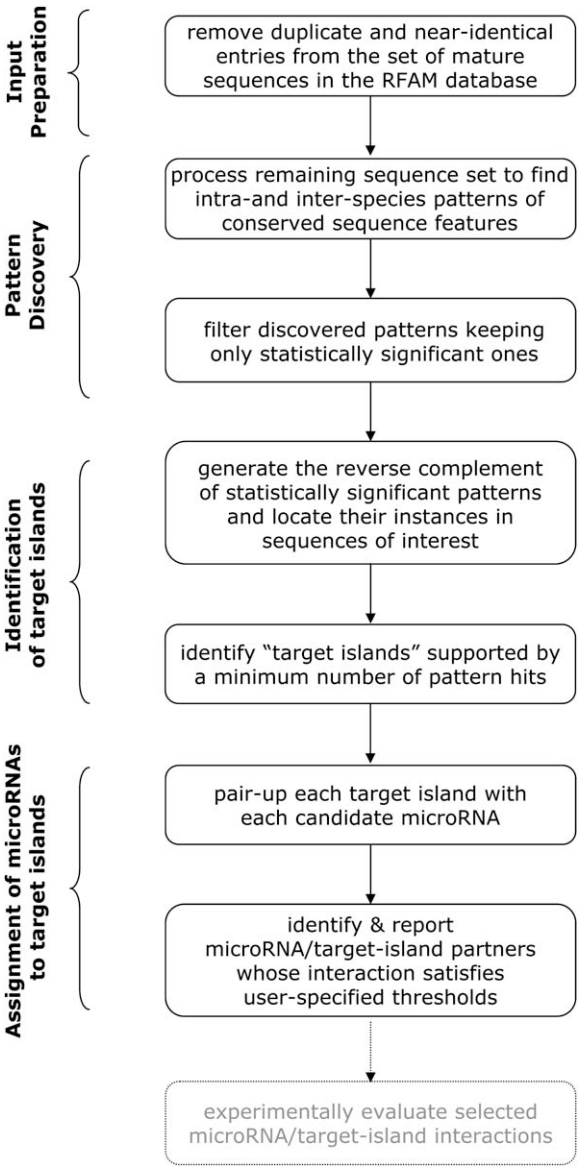


Figure 1. Flowchart Showing the Various Steps of the Method

a second-order Markov chain and used it to estimate each pattern's statistical significance (see Supplemental Data) discarding patterns with estimated log-probability ≥ -38. At the end of this stage, 233,554 mature-microRNA patterns remained.

Identification of "Target Islands"

The key idea behind *rna22* is the use of a redundant yet flexible representation of the knowledge contained in the training set. The latter captures available knowledge in the form of mature microRNA sequences. Through the pattern discovery step, we identify salient, conserved sequence features that we represent using patterns: as a collection, these patterns capture the original body of knowledge but in a different and, importantly, redundant manner: sequence segments from the original input will now be represented by multiple patterns each of which appears, by design, in two or more sequences.

Because the patterns we use are statistically significant, we can treat them as "predicates." If a sequence *T* contains one or more of these patterns, then *T* is likely a member of the same class as the training set sequences from which the patterns were derived – the higher the number of patterns present in *T*, the higher the degree of certainty in this membership. The underpinnings of our method have their origins in the "guilty-by-association" approach that was introduced more than 20 years ago (Doolittle et al., 1983) and gave birth to the field of nucleotide and amino acid sequence comparison. During the last decade, guilty-by-association schemes have been applied with success to many, diverse problems from computational biology (Darzentas et al., 2005; Ettwiller et al., 2003; Murphy et al., 2003; Neduva et al., 2005; Rigoutsos et al., 2002, 2003; Shibuya and Rigoutsos, 2002).

Mature microRNAs bind to 3′UTR targets through hybridization of complementary base pairs. Since our pattern collection captures sequence features of mature microRNAs that are conserved but not necessarily contiguous, it follows that the *reverse complement* of such patterns will permit us to locate conserved sequence elements in the untranslated regions of interest and, thus, putative microRNA binding sites. The dereferenced positions in the reverse complement of each mature microRNA pattern point to salient features shared by the UTR sequence at hand and the reverse complement of known microRNA sequences. The high number of patterns clustering around specific UTR locations in conjunction with the patterns' statistical significance and the "guilty-by-association" approach allow us to associate such "hot spots" with putative microRNA binding sites. It is important to realize that we can identify these hot spots simply by using the multiple patterns from our collection and *without* reference to any specific microRNA. As one might expect, regions that do *not* correspond to binding sites receive a much smaller number of hits allowing us to differentiate between background and *bona fide* binding sites (see Supplemental Data). As an example, [AT][CG].TTTTT [CG]G..[AT][AT][AT]G[CG].CTT is a mature microRNA pattern contained in our collection and AAG.[CG]C[AT][AT] [AT]..C[CG]AAAAA.[CG][AT] is its reverse complement.

We use the term "*target island*" to refer to any hot spot where the reverse complement of mature microRNA patterns aggregate. A pattern's instance contributes a vote of "+1" to all the UTR locations that the instance spans: all sequence regions comprising contiguous blocks of locations receiving $\geq$30 votes are treated as putative microRNA binding sites. It may happen that consecutive locations with above-threshold support span a region *R* with length $\ell_R < 22$ nucleotides: to ensure that we capture the remaining 22-ℓ_R positions of the site whether they lie to the left or the right of *R*, we report a 36-nucleotide segment centered on *R*.

The identification of target islands effectively focuses the algorithm's subsequent steps to only those regions that receive support by the reverse complement of many mature microRNA patterns. The step uses microRNA sequence features to discard UTR regions that are not likely to be microRNA binding sites. This is also a key element behind *rna22*'s noise resilience and we revisit it below in the context of extreme distributions.

Associating MicroRNAs with Target Islands

As we will see, the typical target island is short in length and a fraction of the original UTR's length. Once we locate a putative microRNA binding site in the form of a target island, we determine the identity of the microRNA(s) that will bind to it in a straightforward manner: we pair each one of the available microRNAs with each generated target island, for all possible relative offsets. Given a putative pair of the form microRNA/island-segment we insert the linker GCGGGGACGC (Stark et al., 2003) between the two sequences, form the heteroduplex "microRNA-linker-island segment" and use the Vienna package (Hofacker et al., 1994) to predict the structure of the duplex and its Gibbs free energy ("folding energy").

M, *G* and *E* are three *user-specified* parameters that control the algorithm's output. *M* is the *minimum* required number of base-pairs between the microRNA and the target (excluding base-pairs in the linker). *G* is the *maximum* allowed number of *unpaired* bases in the seed region. *E* is the algebraically maximum allowed free energy, in Kcal/mol. Note that *rna22* imposes no restrictions on the number of G:U pairs that can appear in the seed-region of a complex. Typical settings are $M = 14$, $G = 1$, $E = -20$ Kcal/mol (the linker contributes approx. -5.6 Kcal/mol to the *E* value).

In *rna22*, the reporting of a binding site is not predicated on whether the corresponding sequence is conserved across species. It is also not predicated on the existence of the reverse complement of a microRNA's seed region in the sequence at hand. Instead, *rna22* recognizes a binding site based on the presence of multiple, distinct, statistically significant patterns that have been discovered by processing known mature microRNA sequences. Not needing to know the identity of the microRNA that targets a binding site permits the delineation of binding sites for

as-yet-unidentified microRNAs (see below for a discussion of the *cel-lsy-6* case).

Extreme Distributions, Spurious Binding Sites and the Importance of Target Islands

We can abstract the problem of computational detection of microRNA targets as follows: "given a UTR sequence S_{UTR} of length L_{UTR}, and a specific microRNA m of length $L_{microRNA}$, identify the location(s) in S_{UTR} where m will bind." If $\overline{m}$ denotes the reverse complement of m, then the target detection problem can be cast as one of "local sequence similarity" detection (Gusfield, 1997), where one seeks to locate one or more matches for $\overline{m}$ in S_{UTR} subject to maximizing a quality measure that rewards matches and penalizes insertions, deletions and G:U pairs.

If m binds somewhere in S_{UTR}, then the short segment corresponding to the binding site will be similar to $\overline{m}$. This is not a sufficient condition: $\overline{m}$ can have a good sequence agreement with a segment of S_{UTR} and predicted to form an energetically stable complex, but this does not necessarily mean that the segment in question is a true binding site for m. Such sequence-based agreements, which we refer to as *spurious matches*, have a nonzero probability and will arise by chance when S_{UTR} is a true UTR that does not contain a binding site for the microRNA at hand, or when S_{UTR} is a randomly generated sequence of nucleotides.

The statistics governing spurious sequence matches have been studied in the context of sequence similarity detection (Karlin and Altschul, 1990) and form the backbone of the BLAST suite of algorithms (Altschul et al., 1990). Summarily, when comparing two random sequences of lengths L_1 and L_2, the number of matches that are expected by chance and have score $\geq$ S can be approximated by a Poisson distribution whose mean value E is given by

$$E = K * L_1 * L_2 * \exp(-\lambda S) \quad (1)$$

Here λ and K are constants that depend on the used scoring system and the database respectively. For the problem at hand, L_2 is the length of $\overline{m}$, i.e., $L_2 = L_{\text{microRNA}} \approx 22$ nucleotides. However, $L_1 = L_{UTR}$ and can vary widely. Even though the original derivation of Equation 1 assumed that a local alignment between the two random sequences contained no gaps, this equation has been shown to also apply in the case where gaps are allowed (Altschul and Gish, 1996) and thus we can use it to analyze the case of microRNA target detection.

Equation 1 effectively captures the behavior of all microRNA-target-detection methods that identify binding sites through comparison of the two sequences of nucleotides that form the putative microRNA/mRNA heteroduplex. The equation can also be used as a proxy for estimating the rate of spurious matches when binding sites are identified by analyzing the structure and Gibbs energy of a putative complex (Rehmsmeier et al., 2004): indeed, the higher the sequence similarity between $\overline{m}$ and the putative binding site the better the predicted structure's quality and the lower its Gibbs energy, and vice versa.

The probability that a microRNA target detection algorithm will report one or more spurious microRNA binding sites is P_0 = 1-exp(-E), with E given from Equation 1. P_0 depends on the length of the UTR in which one seeks matches: the longer the UTR, the higher the likelihood that one or more spurious binding sites will be reported for the microRNA at hand. P_0 also depends on the value of the score S of the sequence match between $\overline{m}$ and the putative binding site and, by extension, on the quality of the putative microRNA/mRNA binding: the fewer the matched base pairs in the putative complex the higher the probability that the complex is spurious in nature. Here, we only examine the impact of a UTR's length on the probability of generating spurious matches. This analysis allows us to evaluate qualitatively all of the current microRNA-target-detection methods: the conclusions are applicable to methods that report "full-length" binding sites as well as to "seed-only" methods.

If a target detection algorithm focuses only on a portion of S_{UTR} while ignoring the rest of it, then the probability that this algorithm will report one or more spurious matches is reduced. Indeed, if the algorithm examines only $(1/\rho)$-th of the original UTR length, then substantial gains can be had with increasing ρ values. Figure S2 shows the factor by which the probability of reporting one or more spurious matches is reduced as a function of the value E from Equation 1, and for different values of ρ; microRNA target detection methods that explore the entire UTR sequence correspond to the $\rho = 1$ curve. The plot shows that if an algorithm seeks microRNA matches in only $(1/\rho)$-th of a given UTR, then the probability of the algorithm reporting one or more spurious matches is reduced by a factor approximately equal to ρ when compared to algorithms that seek matches in the full-length UTR. These gains persist for almost four orders of magnitude of E's values, i.e., across a very wide range of real-world scenarios.

Equation 1 and the preceding discussion show the importance of the target island idea. *Rna22* effectively replaces the original UTR sequence by a set of nonoverlapping, small size fragments ("target islands") wherein it seeks microRNA targets. The selection of targets islands is guided by the reverse complement of microRNA-derived patterns. This in turn ensures that *rna22* selects all relevant locations in the UTR at hand before determining whether any of them is targeted by the given microRNA m; we discuss this in more detail in the next section.

Three more questions remain. How many target islands will *rna22* extract from a given UTR? What is the average length of these target islands? Does *rna22* determine the target islands "correctly?"

Figure 2A shows the probability density function for the number of target islands that *rna22* derives from the 25,589 3′UTRs in the processed ENSEMBL release of the human genome. Figure S3 shows the distribution of target island lengths.

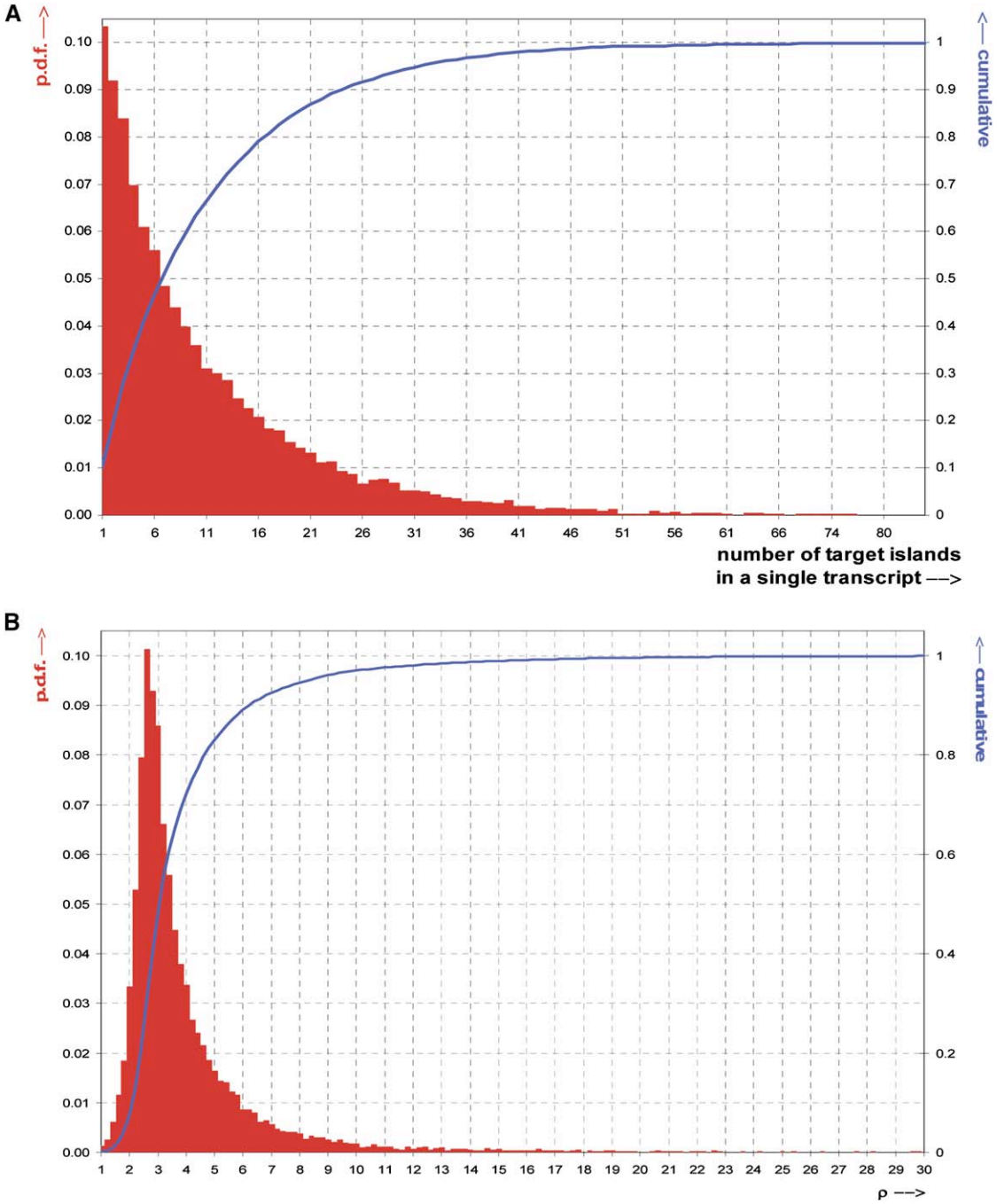


Figure 2. Focusing on a Few, Short Target Islands Instead of the Complete UTR Sequence Improves the Noise Characteristics of MicroRNA Target Identification

(A) Probability density function and cumulative for the number of target islands that *rna22* generates from human 3′UTRs.

(B) Probability density function and cumulative for the effective 3′UTR length reduction ρ by *rna22* (see also text).

In earlier work, Equation 1 was found to be conservative for *short* sequences: it generates too large values for E and a correction is required to account for the so-called "edge effects" (Altschul and Gish, 1996). This correction replaces the original lengths by their "effective lengths" that result if we subtract the length $l_{typical}$ of a typical optimal subalignment; in the case of microRNA target detection $l_{typical}$ is ~20 nucleotides. For target detection algorithms that process full-length UTRs such a correction has negligible impact (Altschul and Gish, 1996) but is


C. elegans
(lin-28) [328,342] 527
(lin-41) [689,710] 1162
(lin-41) [738,757] 1162
(cog-1) [267,285] 394
D. melanogaster
(hid) [874,896] 2269
(hid) [1711,1731] 2269
(hairy) [441,465] 816
(grim) [63,86] 964
(sickle) [34,58] 945
(bagpipe) [87,109] 165
H. sapiens
(hoxb8) [441, 432] 856
(mtpn) [3126,3141] 3165
(lin-28) [890,913] 2751
(mapk-14) [652,669] 2311
(fbxwib) [2393,2415] 2575
(laminin) [210,223] 1504
(dmtf-1) [130,146] 1187
(clock) [215,233] 2922
(cgi-38) [295,311] 363
(smc-1) [74,91] 5945
(bm3b) [103,109] 1663
(bm3b) [464,470] 1663
(enx-1) [60,66] 263
(enx-1) [115,121] 263
(bdnf) [221,227] 2925
(bdnf) [1323,1329] 2925
(g6pd) [98,104] 609
(g6pd) [434,440] 609
(smad) [47,54] 141
(smad) [104,110] 141
(nmyc) [495,501] 904
(sdf-1) [1387,1394] 1570
(sdf-1) [1452,1459] 1570
(mcsf-1) [800,807] 901
M. musculus
(limk1) [180,201] 1170

necessary in the case of *rna22* where the length of most target islands is small.

Replacing a full-length UTR by the corresponding target islands and correcting the islands' lengths to account for "edge effects" reduces the value of L_1 in Equation 1 from L_{UTR} to L_{UTR}^{new} leading to a length reduction by $\rho_{effective} = L_{UTR}/L_{UTR}^{new} \geq 1$. We computed $\rho_{effective}$ for the 25,589 human 3′UTRs in our database and show the corresponding probability density function and the cumulative in Figure 2B. For a large portion of these sequences, the reduction $\rho_{effective}$ of the UTR length is by a factor of 3 or higher. We also computed $\rho_{effective}$ for all worm, fruit fly, and mouse 3′UTRs and the resulting distribution of values is similar (data not shown). Figure S2 shows this length reduction directly translating into a very substantial decrease in the probability of *rna22* reporting spurious matches compared to target detection methods that seek targets by processing full length UTRs. We will see below that, in practice, the generation of target islands has an even more pronounced beneficial effect and that the resulting rate of false positives is low.

An important corollary arises from this analysis. For the four model genomes, namely worm, fruit fly, mouse and human, the average 3′UTR (and 5′UTR) length increases monotonically from worm to human (Pesole et al., 2000). All else being equal, this observation in connection with Equation 1 implies that, on average, any algorithm that reports microRNA binding sites by examining full-length UTR sequences will generate spurious matches with a probability that increases monotonically from worm, to fruit fly, to mouse, to human. In other words, a target detection method that is designed specifically for the fruit fly will report fewer spurious matches, on average, than a method that is designed specifically for the human genome.

Finally, we can see that filtering putative microRNA binding sites by requiring that they be conserved across species can decrease the number of expected spurious matches if and only if the length of the resulting orthologous segment is shorter than the length of the original UTR that contains it. The requirement of cross-species conservation for a binding site might improve a prediction algorithm's noise characteristics but at the cost of potentially discarding true binding sites if the latter are not conserved in the considered species.

In the next section, we examine how well *rna22* captures binding sites that may be present in a given UTR.

Method Validation: Predicting the MicroRNA Binding Sites of Previously Reported Heteroduplexes

To date, only a relatively small number of microRNA binding sites have been validated experimentally in animals, and they come from a handful of species. We evaluated the ability of *rna22* to correctly predict microRNA binding sites by determining how well it identified the sites for the experimentally supported heteroduplexes that have been published to date. This is a nontrivial task if one considers that we trained *rna22* using a January 2004 instance of the RFAM database: consequently, any binding sites that *rna22* predicts and which appeared in the literature after January 2004, are tantamount to correct de novo predictions.

Figure 3 summarizes the results of this test. In each case, we show the plot of pattern support as a function of position within the corresponding 3′UTR. If *rna22* correctly identified a binding site, the area under the support curve is shown in light green, otherwise it is shown in pink. For clarity, we only show the neighborhood around each reported binding site. Each plot title lists the name of the gene, the extent of the reported site and the length of the 3′UTR in this order. We only examined binding sites that were within the extent of the genes' known 3′UTRs. Recall that during target island detection, if the region that exceeds threshold is shorter than 22 nucleotides in length, then we report a 36-nucleotide segment that is centered on the original region (one such example is the case of the *clock* binding site). *Rna22* successfully identifies 81%, or 17 out of the 21 full-length binding sites, i.e., sites with base-pairing that extends beyond the microRNA's seed or nucleus region. Later in the discussion, we revisit this topic and re-plot this graph after training *rna22* with a recent RFAM release.

A notable result is the correct identification of the full extent of *cel-lsy-6*'s binding site in the 3′UTR of the *cog-1* gene from *C. elegans* (Johnston and Hobert, 2003). The importance of the result stems from the fact that *cel-lsy-6* is *not* contained in the January 2004 instance of the RFAM release that we used for training and shares no sequence similarities with any of the microRNAs contained in that release. In other words, the training set contains no explicit information about *cel-lsy-6*, yet *rna22* is able to extrapolate from the available data and correctly identify this binding site.

Method Validation: Identifying the Correct MicroRNA for Previously Reported Heteroduplexes

Once we have identified all target islands in a given UTR, finding the corresponding microRNA that will bind to it amounts to forming complexes between the islands and each candidate microRNA and reporting the microRNA that satisfies the user-specified *M*, *G* and *E* parameters (see above). We applied this process to the 17 previously reported, full-length binding sites: in all 17 cases, *rna22* correctly identified the originally reported microRNA as the one binding to the found site.

Figure 3. Ability of the Method to Identify Previously Reported Binding Sites Using as Training Set a Release of the RFAM Database from January 2004

See text for color convention. The data are from Brennecke et al., 2003; Johnston and Hobert, 2003; Kiriakidou et al., 2004; Lewis et al., 2003; Moss et al., 1997; Poy et al., 2004; Reinhart et al., 2000; Schratt et al., 2006; Stark et al., 2003; and Yekta et al., 2004.

Method Validation: Experimental Support for *Novel* Predictions Made by *rna22*

We used luciferase-reporter assays to test binding sites predicted by *rna22*. Each predicted microRNA binding site was inserted *as a single copy* directly downstream of a *Renilla* luciferase open reading frame (ORF). Examining one predicted target at a time is an important component of our stringent evaluation approach as any reduction in luciferase activity can be attributed to a single source.

We compared the relative luciferase activity of the control transfection (scrambled RNA oligo or empty plasmid vector – represented as 100%) to the activity when the cognate microRNA was added. A sequence antisense to each targeting microRNA formed a positive control. None of the three microRNAs with which we worked was predicted to target the sequence that is antisense to miR-21 so we used this antisense as one of the negative controls. In the Supplemental Data, we list the observed standard deviation about the mean (100%) of luciferase activity for negative controls: these observations served as guidance for requiring that luciferase repression of a predicted target be by *a minimum of* 30% before it can be reported. To ensure repeatability of the tests, and for each one of the predictions that we tested, we carried out the assay three times and with four culture-replicates and recorded the average and standard deviation of relative luciferase activity in each case.

We worked with three murine microRNAs: mmu-miR-375, mmu-miR-296 and mmu-miR-134. MiR-375 was included because its human homolog was characterized and shown to regulate insulin secretion by binding to myotrophin (Poy et al., 2004). The two other microRNAs, miR-296 and miR-134, were included because they are upregulated during embryonic stem cell differentiation.

Requiring a minimum of $M = 14$ matching base pairs between the microRNA and a target, at most $G = 1$ unpaired bases in the seed region, and binding energies [$E = -22$ Kcal/mol for the microRNA/mRNA complex, we predicted 2292, 271 and 2318 targets for miR-375, miR-296 and miR-134 respectively. Since we could not test all these predictions, we randomly subselected among them and tested 44 predicted targets for miR-375, 24 for miR-296, and 158 for miR-134.

Figure 4 shows the results of these assays separately for each microRNA. Luciferase activity was suppressed by at least 30% for 168 out of the 226 tested predictions; in fact, for more than half of the predictions we tested suppression ranged between 40% and 80%. The Supplemental Data lists the ENSEMBL ids and sequences for the 226 tested predictions.

Finally, we note that the recently reported miR-134 binding site in *limk1* (Schratt et al., 2006) is among the set of *rna22*'s predictions but had not been included in the randomly selected set with which we experimented. In subsequent experiments, we confirmed the *limk1* binding site using our reporter assay and found the relative luciferase activity in the presence of miR-134 to be 38.1 ± 6.6%, i.e., miR-134 repressed its target by ~60%.

Method Validation: False Positive Rate, Sensitivity, and Resilience in the Presence of Random Sequences

We estimate *rna22*'s false positive rate in three ways. First, we compute it as the ratio of the number of tested predictions that are repressed by *less* than 30% over the total number of tests. This is equal to 100 − (27 + 12 + 129) / (44 + 24 + 158) = 100 − 168 / 226 = 25.7%.

The second method is more involved. Recall that each of the 168 tested targets for which luciferase repression ≥30% was the unique prediction made by *rna22* for the tested microRNA and the corresponding 3′UTR. As such, these 168 3′UTRs represent a ground truth set in that each of them contains exactly one validated binding site for the corresponding microRNA (miR-375, miR-296, or miR-134). Ideally, if a target detection algorithm were presented with *shuffled* instances of these 3′UTRs, it should report no binding sites for miR-375, miR-296 and miR-134 respectively. Any sites that would be reported on a random input could be used to estimate the false positive rate. *Rna22* reports 4 binding sites in the shuffled instances of the 27 3′UTRs that contain the validated targets for miR-375. Moreover, after processing shuffled instances of the 12 3′UTRs containing the validated miR-296 targets and of the 129 3′UTRs containing the validated miR-134 targets, it reports 0 and 28 binding sites respectively. This is a total of 32 erroneously reported binding sites on these 168 randomized 3′UTRs, or a false positive rate of 32/168 = 19.0%.

The third method takes into account Equation 1 and the relationship between UTR length and the number of spurious binding sites. In our opinion, this is a more realistic measure of an algorithm's performance and we define it as "the average number of binding sites that the algorithm reports per 10,000 nucleotides of randomized input." Since edge effects can be important, one should not attempt to generate estimates of this measure by generating random strings of arbitrary length; rather, this number should be estimated from shuffled instances of 3′UTRs that are known to contain binding sites. The 3′UTRs in which our 168 validated predictions reside total 313,057 nucleotides. Since *rna22* reports 32 spurious binding sites after processing shuffled instances of these sequences we conclude that its performance is "1.0 spurious binding sites for every 10,000 nucleotides of randomized sequence processed."

We conclude by estimating *rna22*'s sensitivity. This measure is more difficult to compute than the false positive rate because sensitivity is linked intimately to the richness of the set that each algorithm uses as its knowledge repository. Moreover, it assumes that all available true positives are obtained using the same protocols and thresholds. First, we approximate *rna22*'s sensitivity using the results of Figure 3: *rna22* identified 81% (17 of 21) of the full-length and 36% (5 of 14) of the seed-only binding sites from among those previously reported, for a total of 22 out of 35 sites, or 63%. To demonstrate the above point, we recomputed the sensitivity of *rna22*

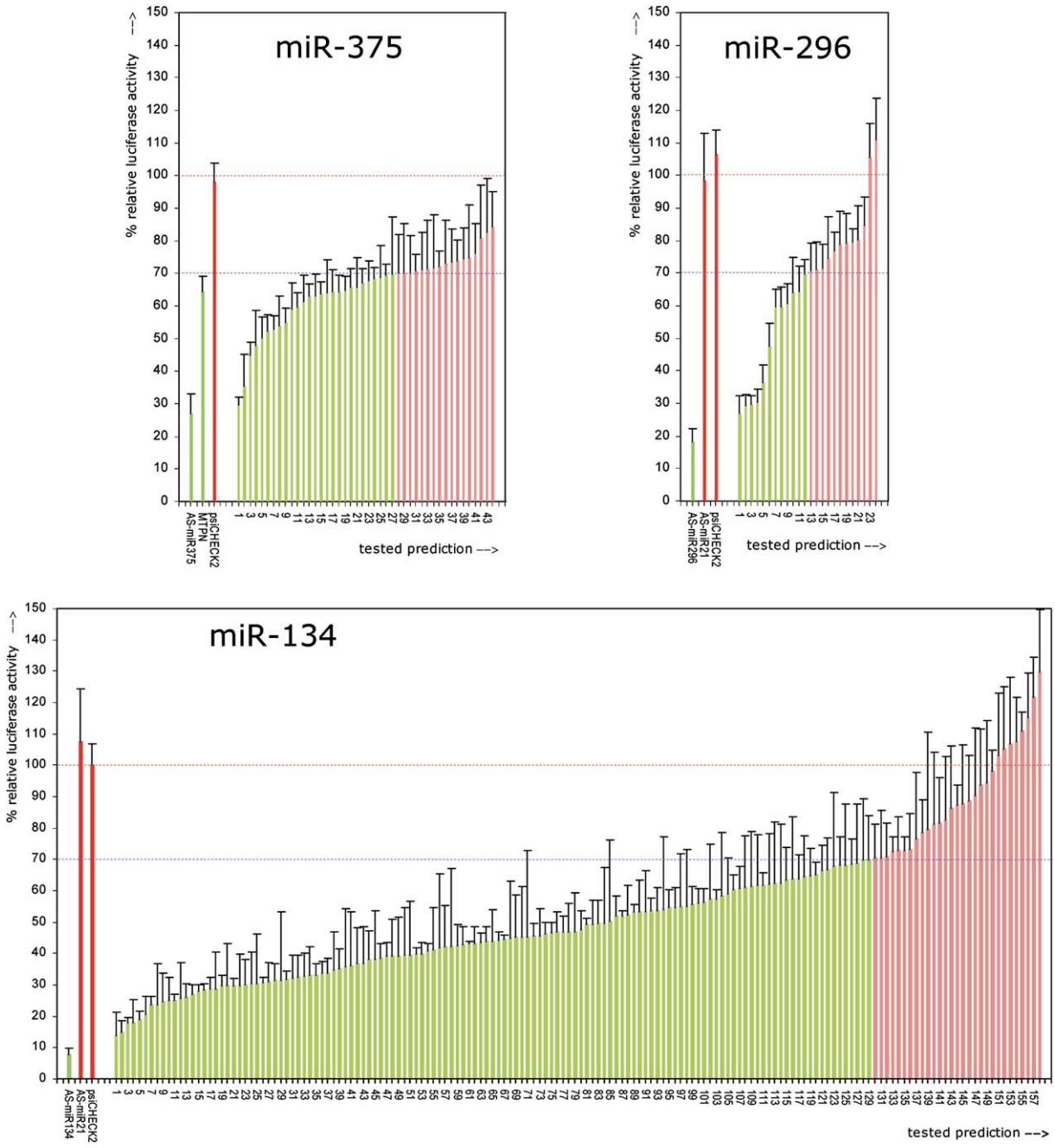


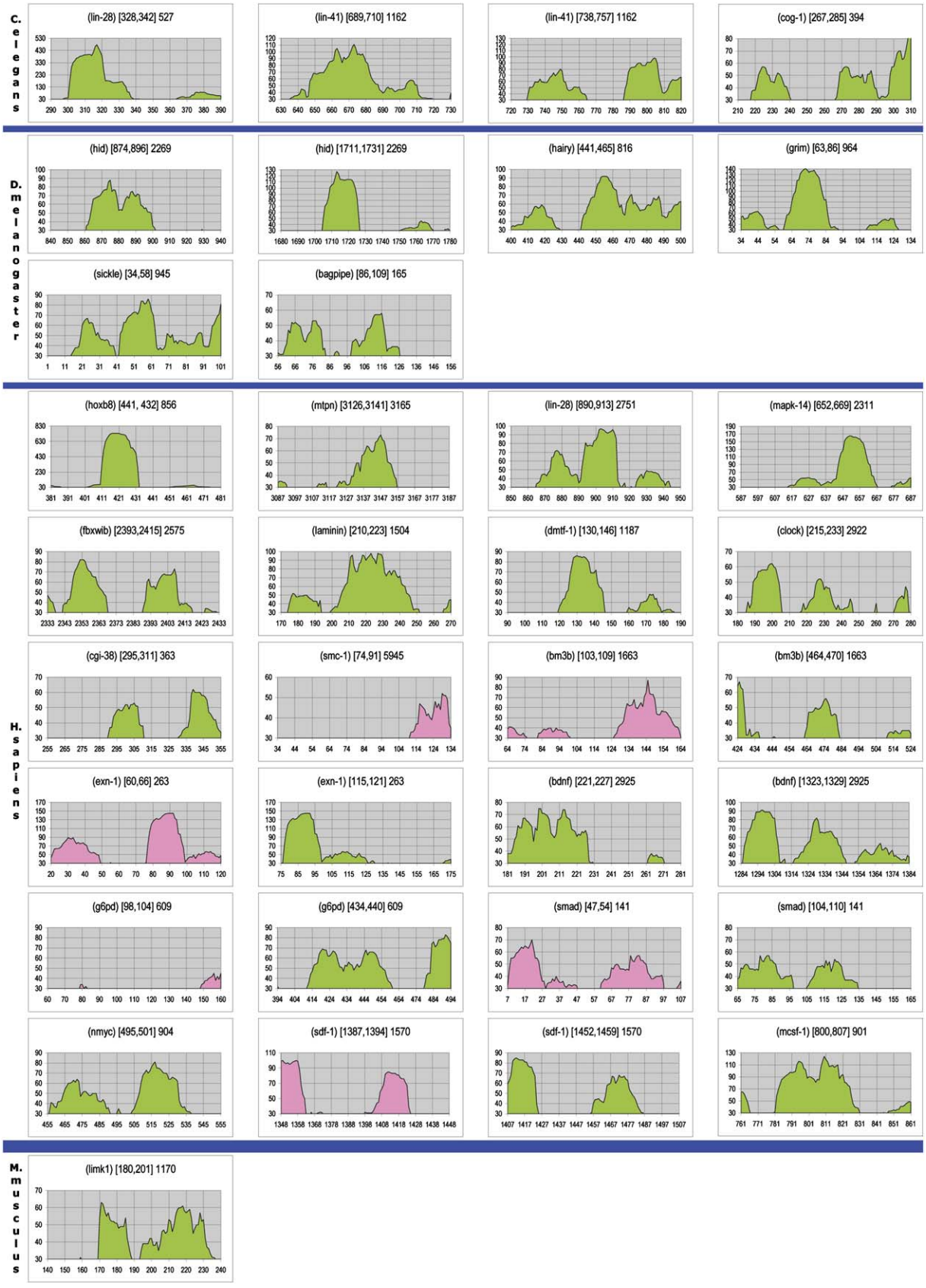
Figure 4. Results of Luciferase-Based Assays of Predicted Targets in 293T Cells for miR-375, miR-296, and miR-134

In all the plots, the y axis shows the relative level of luciferase expression, whereas points along the x axis correspond to the various tested predictions. See Supplemental Data for ENSEMBL identifiers and tested target sequences. Luciferase vector without MRE is shown as psiCHECK2. Error bars, SD for n = 12 repetitions. AS-miRXX indicates the sequence that is anti-sense to miR-XX and XX is one of 375, 296, 134, 21. See also text.

after retraining with a more recent snapshot of RFAM from December 2005. Figure 5 shows the new plots: *rna22* now identifies 95% (20 of 21) of the full-length and 64% (9 of 14) of the seed-only binding sites, or a total of 29 out of 35 sites which corresponds to a sensitivity of 83%.

New Insights on the Rules Governing MicroRNA/Target Recognition

Our reporter assays permit several observations regarding the rules that govern target recognition by microRNAs. Some of our mouse-derived results differ from earlier reports that were obtained in the fruit fly (Brennecke et al.,


C. elegans
(lin-28) [328,342] 527
(lin-41) [689,710] 1162
(lin-41) [738,757] 1162
(cog-1) [267,285] 394
D. melanogaster
(hid) [874,896] 2269
(hid) [1711,1731] 2269
(hairy) [441,465] 816
(grim) [63,86] 964
(sickle) [34,58] 945
(bagpipe) [86,109] 165
H. sapiens
(hoxb8) [441, 432] 856
(mtpn) [3126,3141] 3165
(lin-28) [890,913] 2751
(mapk-14) [652,669] 2311
(fbxwib) [2393,2415] 2575
(laminin) [210,223] 1504
(dmtf-1) [130,146] 1187
(clock) [215,233] 2922
(cgi-38) [295,311] 363
(smc-1) [74,91] 5945
(bm3b) [103,109] 1663
(bm3b) [464,470] 1663
(exn-1) [60,66] 263
(exn-1) [115,121] 263
(bdnf) [221,227] 2925
(bdnf) [1323,1329] 2925
(g6pd) [98,104] 609
(g6pd) [434,440] 609
(smad) [47,54] 141
(smad) [104,110] 141
(nmyc) [495,501] 904
(sdf-1) [1387,1394] 1570
(sdf-1) [1452,1459] 1570
(mcsf-1) [800,807] 901
M. musculus
(limk1) [180,201] 1170

2005) providing yet another testimony of how complex the problem at hand is. Analysis of our 168 validated targets showed that:

Extensive, Bulge-Free Base-Pairing to the 5′ End of the MicroRNA May Not Always Lead to Substantial Repression

Tested miR-375 targets #32, 34, 40, 42, and 43 (see Figure 4 and Supplemental Data) are predicted to form complexes with 9, 10, 11, 10, and 11 base pairs respectively that are uninterrupted, bulge-free and begin at position 2 from the microRNA's 5′ end; despite these extensive base pairings, the observed repression was less than 30%.

Multiple G:U Pairs Simultaneously Present with One or More Single-Nucleotide Bulges in the Seed Region Can Still Lead to Substantial Repression

Tested miR-134 target #79 (Figure 4 and Supplemental Data) is predicted to form a complex with 2 G:U pairs and a single nucleotide bulge in the seed region yet it is repressed by 53%. Tested miR-134 targets #73, 125, and 128 are predicted to form complexes with 3 G:U pairs and a single nucleotide bulge in the seed region yet are repressed by ~54%, 32% and 30% respectively. Target #94 forms a complex with 5 G:U pairs and a single-nucleotide bulge in the seed region yet is repressed by 46%. And finally, the complex for target #97 is repressed by 45% even though it contains 4 G:U pairs and two single-nucleotide bulges in the seed region.

Asymmetric, Single-Nucleotide Seed-Region Bulges on the MicroRNA Side Do Not Necessarily Abolish Repression

The complexes for tested miR-134 targets #131, 105, 128, 41, 65, and 127 (Figure 4 and Supplemental Data) contain no bulges in the region of the target opposite to the microRNA seed but have a single nucleotide bulge at positions 3, 5, 6, 7, 7, and 7 respectively of the microRNA seed. Despite these asymmetric bulges on the microRNA side, the corresponding targets are repressed by between ~30% and 64%.

Genome-Wide Estimates of MicroRNA Binding Sites in 3′UTRs

We analyzed the 3′UTRs of *C.elegans*, *D. melanogaster, M. musculus* and *H. sapiens* and estimated the number of microRNA binding sites they contain. Table 1A summarizes the findings: depending on the genome, between 74% and 92% of the known 3′UTRs contain one or more target islands (each of which corresponds to at least one putative binding site).

Extending the Method to the Discovery of MicroRNA Precursors

We also adapted the key idea underlying our microRNA target detection method to the problem of discovering microRNA precursors. We now make use of patterns derived from microRNA *precursors* to query an organism's genomic sequence and determine regions where these patterns aggregate: these regions are predicted to be putative microRNA precursors. We can find the corresponding mature microRNA(s) in these putative precursors simply by locating instances of the mature-microRNA-patterns and identifying which regions accumulate high-hit counts (see Supplemental Data).

We first used as our training set the 719 microRNA *precursor* sequences contained in Release 3.0 of RFAM (01/2004). After removing identical and near-duplicate entries, we obtained a nonredundant set of 530 sequences that we then processed with Teiresias. After discarding statistically insignificant patterns, a total of 192,240 microRNA-precursor patterns remained. Processing the intergenic and intronic regions of the four organisms of interest shows that we can identify de novo more than 3/4 of the microRNA precursors from the January 2004 training set. If we now repeat the above using as our training set the precursor sequences contained in the December 2005 instance of RFAM, the sensitivity rises to ~93% (see Table S4).

In addition to discovering known precursors, the method also predicts many previously unreported, putative precursors. Table 2 lists the number of predicted precursors for four different genomes and for two folding-energy cutoffs: −25 Kcal/mol and −18 Kcal/mol. Even at the more stringent threshold of −25 Kcal/mol, our precursor estimates for all four genomes are substantially higher than what has been suggested previously (Berezikov et al., 2006). The error rate for this precursor discovery scheme is estimated to be between 1% and 2% (see Supplemental Data). The high number of predicted human precursors is in concordance with the results of a recent independent analysis of human intergenic and intronic regions (Rigoutsos et al., 2006).

DISCUSSION

We presented a pattern-based methodology for the identification of microRNA binding sites and the corresponding heteroduplexes. We also extended it to the discovery of microRNA precursors. We demonstrated the method's power by providing computational and experimental evidence and applied it to the analysis of several genomes. Based on our findings, and in addition to the results presented above, we can observe the following:

Insulin Secretion in Murinae

Validated target #2 of miR-375 (see Supplemental Data for sequence information) is in the 3′UTR of Kv2, a member of the voltage-dependent K+ channel family. In addition,

Figure 5. Improvement in *rna22*'s Ability to Identify Previously Reported Binding Sites after Training with the December 2005 Release of the RFAM Database

The color coding convention and shown data are as in Figure 3. See also text.

Table 1. Summary of *Rna22*'s Predictions for Four Model Genomes

A

Genome	Number of Processed 3′UTRs	Number of 3′UTRs Containing One or More "Target Islands" (% Processed 3′UTRs)	Number of Nucleotides in Processed 3′UTRs	Number of "Target Islands" in Processed 3′UTRs
C. elegans	13,186	9752 (73.9%)	3,048,704	27,700
D. melanogaster	14,965	13,104 (87.6%)	6,671,035	63,918
M. musculus	20,257	18,597 (91.8%)	18,058,224	180,157
H. sapiens	25,589	23,616 (92.3%)	25,597,040	243,211

B

Genome	Number of Processed 5′UTRs	Number of 5′UTRs Containing One or More "Target Islands" (% Processed 5′UTRs)	Number of Nucleotides in Processed 5′UTRs	Number of "Target Islands" in Processed 5′UTRs
C. elegans	11,713	3654 (31.2%)	797,941	7085
D. melanogaster	15,461	12,139 (32.7%)	4,129,409	37,078
M. musculus	19,978	10,298 (51.5%)	4,398,970	31,967
H. sapiens	25,042	13,350 (53.3%)	6,947,437	46,007

C

Genome	Number of Processed CDSs	Number of CDSs Containing One or More "Target Islands" (% Processed CDSs)	Number of Nucleotides in Processed CDSs	Number of "Target Islands" in Processed CDSs
C. elegans	25,811	23,515 (91.1%)	34,476,529	362,110
D. melanogaster	19,177	19,059 (99.4%)	32,199,294	270,617
M. musculus	31,535	31,345 (99.4%)	42,926,064	420,238
H. sapiens	33,869	33,545 (99.0%)	50,737,171	476,677

(A) Results from the analysis of 3′UTRs.
(B) Results from the analysis of 5′UTRs.
(C) Results from the analysis of CDSs.

validated target #14 is in the 3′UTR of a GLP-2 receptor. Both of these targets are linked to insulin secretion (Kawai et al., 1995; MacDonald et al., 2001) raising the possibility that in mice/rats miR-375 may modulate the latter by acting on myotrophin and additional targets of the pathway.

A Single MicroRNA Can Have Numerous Targets

Most striking among the assays are those obtained for miR-134 and miR-375, where in 81.7% and 61.4% of the tested cases respectively repression was by at least 30%. If this ratio of success, which we observed by testing randomly selected targets, is representative of the situation then it follows that a large portion of the predicted miR-134 and miR-375 targets are likely true. In other words, microRNA miR-134 may have more than 129/158*2318 = ~1,890 targets, and miR-375 may have more than 27/44*2292 = ~1,400 targets. We conjecture that these two microRNAs are not an exception, and that many more microRNAs have numerous targets. This hypothesis is in congruence with the average number of transcripts that *rna22* estimates are targeted by a single microRNA: focusing on 3′UTR targets alone, the numbers range from several tens of targets in the worm to about a thousand in the human genome (see Table S3).

Revisiting the Number and Location of MicroRNA Binding Sites

Since the presented method is not biased in any way in favor of 3′UTRs (its starting point is microRNA sequences and not 3′UTR sequences), we also applied it to the analysis of the 5′UTRs and CDSs of the four genomes under consideration. Parts (B) and (C) of Table 1 summarize our findings. Depending on the genome, between 31% and 53% of the known 5′UTR sequences are predicted to contain one or more targets. For CDSs, almost every amino acid coding sequence is predicted to contain one or more targets, in an apparently genome-independent manner. It is interesting that in all cases (5′UTR, CDS,

Table 2. *Rna22*'s Estimates of the Number of MicroRNA Precursors for the Worm, Fruit Fly, Mouse, and Human Genomes

Genome	Number of MicroRNA Precursors Contained in the Used Training Set	Number of MicroRNA Precursors that Are in the Training Set and Can Be Detected by *rna22*	Total Number of MicroRNA Precursors Detected by *rna22* Including Already Known Ones ≤ -25 Kcal/mol (≤ -18 Kcal/mol)	Estimated Error when Predicting MicroRNA Precursors ≤ -25 Kcal/mol (≤ -18 Kcal/mol)
C. elegans	106	78 (73.6%)	359 (745)	$\leq$1% ($\leq$2%)
D. melanogaster	78	62 (79.5%)	654 (1,236)	$\leq$1% ($\leq$2%)
M. musculus	202	165 (81.7%)	>25,000 (>44,000)	$\leq$1% ($\leq$2%)
H. sapiens	176	154 (87.5%)	>25,000 (>55,000)	$\leq$1% ($\leq$2%)

Results are reported for two folding energy cutoffs: -25 Kcal/mol and -18 Kcal/mol.

3′UTR) the number of discovered target islands is roughly 1/100 of the number of examined nucleotides. Even though we provided evidence for the method's low rate of false positives in the case of 3′UTRs, it is not clear that we can extrapolate and assume that the same error rate also applies to *rna22*'s predictions in 5′UTRs and CDSs. Nonetheless, even if we dismiss a considerable fraction of the predicted binding sites in 5′UTRs and CDSs as erroneous, the large cardinalities raise the distinct possibility that fairly extensive microRNA regulation may be effected through the 5′UTRs and CDSs of gene transcripts in animals, in addition to 3′UTRs.

Revisiting the Number of MicroRNA Precursors

Analysis of the four model organisms indicates that the number of endogenously encoded microRNA precursors may in fact be substantially higher than currently hypothesized, in all four studied genomes (Table 2). The method's estimated low false-positive rates further support this conjecture. We emphasize that, for each predicted precursor, *rna22* also reports the mature microRNA(s) contained therein, in direct contrast to other methods that often cannot predict the location of mature microRNAs (for a discussion, see Berezikov et al., 2006) (see Supplemental Data for an example).

EXPERIMENTAL PROCEDURES

Cell Culture

HEK 293T/17 (ATCC: CRL-11268) cells were cultured in Dulbecco's modified Eagle's medium (DMEM; Gibco Life Technology, MD, USA, http://www.invitrogen.com) supplemented with 10% heat-inactivated fetal bovine serum (FBS; Gibco) and penicillin/streptomycin (Gibco), maintained at 37°C with 5% CO_2.

Synthetic MicroRNA Oligos

Pre-miR™ microRNA-134 precursor (134 MM) and the scrambled (Scr) RNA oligomer (AGACUAGCGGUAUCUUUAUCCC) were from Ambion, TX, USA, http://www.ambion.com.

MiR-375/MiR-296/MiR-21 Overexpression Vector Construction

To generate the overexpression vector for mmu-miR-375, a 500bp fragment was amplified by PCR from mouse genomic DNA using the Expand High Fidelity system (Roche Diagnostics, Germany, http://www.roche-applied-science.com) and inserted into a modified pIRES-EGFP vector (EcoRI and BamHI sites; Clontech, CA, USA, http://www.clontech.com). To generate the mmu-miR-296 and mmu-miR-21 overexpression vectors, 500bp fragments were amplified by PCR from mouse genomic DNA using the Expand High Fidelity system (Roche) and inserted into the pLL3.7 lentiviral vector (*Xho* I and *Hpa* I sites; a kind gift from the Center for Cancer Research, MIT).

Luciferase Reporter Construct and Target Validation Assay

The predicted microRNA binding sites (= microRNA-response-element or MRE) were synthesized as sense and antisense oligomers, annealed and cloned into psiCHECK-2 (*Xho* I and *Not* I sites, Promega, WI, USA http://www.promega.com), directly 3′-downstream of *Renilla* Luciferase (MRE-*R*Luc). 293T cells were seeded 24 hr before transfection at a density of 5×10^4 cells/well in 96-well plates. In the target validation of miR-375 and miR-296, 120 ng of overexpression vector or empty vector were cotransfected with 2 ng of the MRE-*R*Luc reporter vector using Lipofectamine 2000 (Invitrogen, CA, USA, http://www.invitrogen.com). In the target validation of miR-134, 12.5 nM of miR-134 MM or Scr oligo were cotransfected with 2 ng of the MRE-*R*Luc vector. Concurrently, additional controls were also performed using unpredicted MRE-*R*Luc (e.g., antisense to miR-21) vs cognate microRNA or predicted MRE-*R*Luc vs noncognate microRNAs (e.g., mmu-miR-21). In all cases, a constitutively expressed Firefly luciferase gene activity in psiCHECK-2 served as a normalisation control for transfection efficiency. 48h posttransfection, Firefly and *Renilla* luciferase activities were measured consecutively with the Dual-Luciferase® Reporter system (Promega) by a luminometer (Centro LB960; Berthold Technologies Gmbh & Co. KG, Germany, http://www.bertholdtech.com). All luciferase assays were repeated a minimum of three times with 4 culture replicates each.

Statistical Analysis

Student's nonpaired t test was used to determine the significance of transfected cells relative to control transfected cells.

Supplemental Information

An implementation of *rna22* is available online at: http://cbcsrv.watson.ibm.com/rna22.html.

Supplemental Data

Supplemental Data include six figures, four tables, Supplemental References, and three Excel spreadsheets and can be found with this article online at http://www.cell.com/cgi/content/full/126/6/1203/DC1/.

ACKNOWLEDGMENTS

We thank John Mattick and Edison Liu for their support and encouragement throughout this project. Special thanks to Sumin Koo and Leng Siew Yap for their help and input. We also thank our colleagues at IBM Research and the GIS for their suggestions and for stimulating discussions. Y.T. and W-L.T. received scholarship funds from A*Star, Singapore. B.L. was partially supported by National Institutes of Health (NIH) Grants DK04763 and AI54973.

Received: April 5, 2006
Revised: June 16, 2006
Accepted: July 26, 2006
Published: September 21, 2006

REFERENCES

Altschul, S.F., and Gish, W. (1996). Local alignment statistics. Methods Enzymol. *266*, 460–480.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Berezikov, E., Cuppen, E., and Plasterk, R.H. (2006). Approaches to microRNA discovery. Nat. Genet. *38* (*Suppl 1*), S2–S7.

Brazma, A., Jonassen, I., Eidhammer, I., and Gilbert, D. (1998). Approaches to the automatic discovery of patterns in biosequences. J. Comput. Biol. *5*, 279–305.

Brennecke, J., Hipfner, D.R., Stark, A., Russell, R.B., and Cohen, S.M. (2003). bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. Cell *113*, 25–36.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. (2005). Principles of microRNA-target recognition. PLoS Biol. *3*, e85 10.1371/journal.pbio.0030085.

Croce, C.M., and Calin, G.A. (2005). miRNAs, cancer, and stem cell division. Cell *122*, 6–7.

Darzentas, N., Rigoutsos, I., and Ouzounis, C.A. (2005). Sensitive detection of sequence similarity using combinatorial pattern discovery: a challenging study of two distantly related protein families. Proteins *61*, 926–937.

Doolittle, R.F., Hunkapiller, M.W., Hood, L.E., Devare, S.G., Robbins, K.C., Aaronson, S.A., and Antoniades, H.N. (1983). Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. Science *221*, 275–277.

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

Ettwiller, L.M., Rung, J., and Birney, E. (2003). Discovering novel cis-regulatory motifs using functional networks. Genome Res. *13*, 883–895.

Filipowicz, W. (2005). RNAi: the nuts and bolts of the RISC machine. Cell *122*, 17–20.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. (2003). Rfam: an RNA family database. Nucleic Acids Res. *31*, 439–441.

Gusfield, D. (1997). Algorithms on strings, trees, and sequences: computer science and computational biology (Cambridge, UK; New York, NY: Cambridge University Press).

Hammond, S.M. (2005). Dicing and slicing: the core machinery of the RNA interference pathway. FEBS Lett. *579*, 5822–5829.

Hannon, G.J. (2002). RNA interference. Nature *418*, 244–251.

Hobert, O. (2004). Common logic of transcription factor and microRNA action. Trends Biochem. Sci. *29*, 462–468.

Hofacker, I.L., Fontana, W., Stadler, P., Bonhoeffer, L.S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. Monatshefte f Chemie. *125*, 167–188.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. (2004). Human MicroRNA targets. PLoS Biol. *2*, e363 10.1371/journal.pbio.0020363.

Johnston, R.J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in Caenorhabditis elegans. Nature *426*, 845–849.

Karlin, S., and Altschul, S.F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA *87*, 2264–2268.

Kawai, K., Yokota, C., Ohashi, S., Watanabe, Y., and Yamashita, K. (1995). Evidence that glucagon stimulates insulin secretion through its own receptor in rats. Diabetologia *38*, 274–276.

Kiriakidou, M., Nelson, P.T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. Genes Dev. *18*, 1165–1178.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. Nat. Genet. *37*, 495–500.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. Science *294*, 853–858.

Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. RNA *9*, 175–179.

Lee, R.C., and Ambros, V. (2001). An extensive class of small RNAs in Caenorhabditis elegans. Science *294*, 862–864.

Lewis, B.P., Burge, C.B., and Bartel, D.P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15–20.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. (2003). Prediction of mammalian microRNA targets. Cell *115*, 787–798.

MacDonald, P.E., Ha, X.F., Wang, J., Smukler, S.R., Sun, A.M., Gaisano, H.Y., Salapatek, A.M., Backx, P.H., and Wheeler, M.B. (2001). Members of the Kv1 and Kv2 voltage-dependent K(+) channel families regulate insulin secretion. Mol. Endocrinol. *15*, 1423–1435.

Mattick, J.S., and Makunin, I.V. (2005). Small regulatory RNAs in mammals. Hum. Mol. Genet. *14*(Spec No 1), R121–R132.

Moss, E.G., Lee, R.C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in C. elegans and is regulated by the lin-4 RNA. Cell *88*, 637–646.

Murphy, E., Rigoutsos, I., Shibuya, T., and Shenk, T.E. (2003). Reevaluation of human cytomegalovirus coding potential. Proc. Natl. Acad. Sci. USA *100*, 13585–13590.

Neduva, V., Linding, R., Su-Angrand, I., Stark, A., de Masi, F., Gibson, T.J., Lewis, J., Serrano, L., and Russell, R.B. (2005). Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol. *3*, e405 10.1371/journal.pbio.0030405.

Pesole, G., Grillo, G., Larizza, A., and Liuni, S. (2000). The untranslated regions of eukaryotic mRNAs: structure, function, evolution and bioinformatic tools for their analysis. Brief. Bioinform. *1*, 236–249.

Poy, M.N., Eliasson, L., Krutzfeldt, J., Kuwajima, S., Ma, X., Macdonald, P.E., Pfeffer, S., Tuschl, T., Rajewsky, N., Rorsman, P., and Stoffel, M. (2004). A pancreatic islet-specific microRNA regulates insulin secretion. Nature *432*, 226–230.

Rajewsky, N. (2006). microRNA target predictions in animals. Nat. Genet. *38* (*Suppl 1*), S8–S13.

Rajewsky, N., and Socci, N.D. (2004). Computational identification of microRNA targets. Dev. Biol. *267*, 529–535.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. (2004). Fast and effective prediction of microRNA/target duplexes. RNA *10*, 1507–1517.

Reinhart, B.J., Slack, F.J., Basson, M., Pasquinelli, A.E., Bettinger, J.C., Rougvie, A.E., Horvitz, H.R., and Ruvkun, G. (2000). The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature *403*, 901–906.

Rigoutsos, I., and Floratos, A. (1998). Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics *14*, 55–67.

Rigoutsos, I., Huynh, T., Floratos, A., Parida, L., and Platt, D. (2002). Dictionary-driven protein annotation. Nucleic Acids Res. *30*, 3901–3916.

Rigoutsos, I., Huynh, T., Miranda, K., Tsirigos, A., McHardy, A., and Platt, D. (2006). Short blocks from the noncoding parts of the human genome have instances within nearly all known genes and relate to biological processes. Proc. Natl. Acad. Sci. USA *103*, 6605–6610.

Rigoutsos, I., Riek, P., Graham, R.M., and Novotny, J. (2003). Structural details (kinks and non-alpha conformations) in transmembrane helices are intrahelically determined and can be predicted by sequence pattern descriptors. Nucleic Acids Res. *31*, 4625–4631.

Schratt, G.M., Tuebing, F., Nigh, E.A., Kane, C.G., Sabatini, M.E., Kiebler, M., and Greenberg, M.E. (2006). A brain-specific microRNA regulates dendritic spine development. Nature *439*, 283–289.

Shibuya, T., and Rigoutsos, I. (2002). Dictionary-driven prokaryotic gene finding. Nucleic Acids Res. *30*, 2710–2725.

Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003). Identification of Drosophila MicroRNA targets. PLoS Biol. *1*, e60 10.1371/journal.pbio.0000060.

Yekta, S., Shih, I.H., and Bartel, D.P. (2004). MicroRNA-directed cleavage of HOXB8 mRNA. Science *304*, 594–596.