BIOINFORMATICS DISCOVERY NOTE

Vol. 27 no. 10 2011, pages 1346–1350
doi:10.1093/bioinformatics/btr149

Sequence analysis

Advance Access publication March 26, 2011

The sufficient minimal set of miRNA seed types

Daniel C. Ellwanger[1], Florian A. Büttner[1,*], Hans-Werner Mewes[1,2] and Volker Stümpflen[1,*]

[1]Institute of Bioinformatics and Systems Biology (MIPS), Helmholtz Zentrum München - German Research Center for Environmental Health, D-85764 Neuherberg and [2]Chair of Genome-oriented Bioinformatics, Technische Universität München, Center of Life and Food Science, D-85350 Freising-Weihenstephan, Germany

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Pairing between the target sequence and the 6–8 nt long seed sequence of the miRNA presents the most important feature for miRNA target site prediction. Novel high-throughput technologies such as Argonaute HITS-CLIP afford meanwhile a detailed study of miRNA:mRNA duplices. These interaction maps enable a first discrimination between functional and non-functional target sites in a bulky fashion. Prediction algorithms apply different seed paradigms to identify miRNA target sites. Therefore, a quantitative assessment of miRNA target site prediction is of major interest.

Results: We identified a set of canonical seed types based on a transcriptome wide analysis of experimentally verified functional target sites. We confirmed the specificity of long seeds but we found that the majority of functional target sites are formed by less specific seeds of only 6 nt indicating a crucial role of this type. A substantial fraction of genuine target sites are non-conserved. Moreover, the majority of functional sites remain uncovered by common prediction methods.

Contact: florian.buettner@helmholtz-muenchen.de; v.stuempflen@helmholtz-muenchen.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on October 18, 2010; revised on March 15, 2011; accepted on March 17, 2011

1 INTRODUCTION

The relation between miRNAs and their targets in higher eukaryotes is part of the highly complex gene regulation network. To unravel the functional specific interactions, the available information on the interaction of the short RNAs as presented by the RISC complex and their mRNA counterparts is insufficient to reliably predict all functional pairs modulating translation and mRNA decay (Bagga *et al.*, 2005; Guo *et al.*, 2010; Lim *et al.*, 2005).

The basic prerequisite for miRNA targeting in metazoans is a short perfect match complemented by imperfect matches in close vicinity. This region is called the seed sequence and is considered to be a 6–8 nt long substring within the first 8 nt at the 5′-end of the miRNA (Lewis *et al.*, 2003). It is regarded to be the most important feature for target recognition by miRNAs in mammalians (Bartel, 2009; Nielsen *et al.*, 2007).

Naturally, merely seeking for short sequence matches yields a plethora of putative target sites with a large fraction of false positives. To dodge *a priori* the majority of false positives *in silico* miRNA target site prediction approaches concentrate on the subset of target sites equipped with long perfect seed matches. In addition, several miRNA targeting determinants beyond the seed have been proposed (Grimson *et al.*, 2007; Hausser *et al.*, 2009; Kertesz *et al.*, 2007) to extract authentic target sites from the set of seed matches. A common strategy to increase specificity is to require conservation of the seed match. But there is evidence that non-conserved miRNA targeting is even more widespread (Baek *et al.*, 2008; Farh *et al.*, 2005).

To date the effect of different types of seed matches has been assessed by means of signal-to-noise ratio (Lewis *et al.*, 2003, 2005), degree of mRNA (Grimson *et al.*, 2007, Nielsen *et al.*, 2007) or protein repression (Baek *et al.*, 2008, Selbach *et al.*, 2008). Based on that, a set of canonical seed types that differ in abundance and intensity of the regulatory effect has been defined (Bartel, 2009). Recent experimental approaches allow for the identification of Argonaute (Ago)-miRNA:mRNA ternary complexes using an *in vivo* cross-linking protocol and subsequent high-throughput sequencing (Chi *et al.*, 2009; Hafner *et al.*, 2010). Chi *et al.* (2009) analyzed miRNA:mRNA interactions in *Mus musculus* neocortex tissue samples and published an interaction map containing a set of verified target sites in the transcriptome of the murine brain.

Complementing previous studies, we determined seed types using functional target sites of the interaction map. We identified a minimal and sufficient set of six seed types. The precise mapping of Ago footprints allowed us to distinguish between miRNA:target and higher resolved miRNA:target site interaction. We quantified the impact of individual seed types on recall and specificity. Additional target site conservation analyses revealed short seeds to be less conserved than long seeds.

2 MATERIALS AND METHODS

Chi *et al.* (2009) provided a miRNA:mRNA interaction map that contains the absolute chromosomal positions of sites full complementary to miRNA seeds (murine genome assembly of 2006). These sites are located almost at the center of an average Ago-mRNA footprint. This is a defined region of mRNA complexed with Ago determined by Ago-mRNA clusters, where Ago bound within 62 nt of cluster peaks ≥ 95% of the time. For each chromosomal coordinate, we determined the longest protein-coding mature mRNA transcript and its corresponding relative position by means of the NCBI reference sequence database (Pruitt *et al.*, 2009). Sites that were located within an intron (4%) or upstream of the 3′UTR (45%) were

*To whom correspondence should be addressed.

removed. Ago HITS-CLIP included 20 miRNAs, whereas 18 of which are broadly conserved [according to (Friedman *et al.*, 2009)]. We proceeded with conserved miRNAs. All mRNA and miRNA data were downloaded from UCSC (Karolchik *et al.*, 2004) and miRBase (Griffiths-Jones, 2010) on October 2010.

Based on the set of conserved miRNA sequences and mRNA 3′ UTR sequences, we determined all sites complementary to a minimum of six contiguous nucleotides beginning at either position one, two or three relative to the 5′-end of the miRNA. Seed matches were classified functional or non-functional by means of their distance to Ago HITS-CLIP sites. To account for all seed start positions, each seed match located within a distance of two nucleotides to an Ago HITS-CLIP site was tagged functional. Since the Ago HITS-CLIP sites were located almost at the center of an average Ago-mRNA footprint, matches located within a distance of 3–31 nt could also be functional. Since the locations of the footprints were not available, an unambiguous classification was not feasible. To avoid false positives, these sites remained unclassified. All seed matches located beyond the footprint (distance >31) were classified as non-functional. Further, two miRNAs whose target sites were not significantly enriched in the footprints were removed from the dataset (Supplementary Table S2). Finally, we got 7342 functional, 64 689 non-functional and 1755 unclassified seed sites. Verifying a required minimum target site length of 6 nt, we determined all 5mer matches. The frequency of seed matches within a footprint (distance ≤ 31) and beyond of it was calculated for each seed match length. Additionally, to support our results we prepared the data of the PAR-CLIP experiment in a quite similar fashion (Hafner *et al.*, 2010) (Supplementary Material).

We defined the background set Ω based on the functional and non-functional sites. A seed match $s \in \Omega$ was distinguished by its start position relative to the miRNA 5′UTR ($1=\alpha$, $2=\beta$, $3=\gamma$) and its length. The outcome of this were 20 match types $S_{p,k}$ for a length k and a start position type p. The distributions of all seed match types were disjoint that is each seed match was graded by the longest possible type. To reduce unnecessary complexity of the seed match type set, we merged iteratively non-significant seed match types with their superset. Due to the hierarchical structure of Ω, we were able to apply a separate-and-conquer algorithm (Supplementary Algorithm S1). First we divided the target sites by their seed match start position. Thus, we got three supersets composed of seed matches of a minimum length of 6 nt containing all seed types: $S^{+}_{\alpha,6}$, $S^{+}_{\beta,6}$, $S^{+}_{\gamma,6}$. These sets were separated into 6mers having a mismatch at their subsequent position ($S_{\alpha,6}$, $S_{\beta,6}$, $S_{\gamma,6}$) and seed matches having a minimum length of 7 nt $S^{+}_{p,7}$. We tested the null hypothesis stating that the distribution of functional and non-functional target sites is independent of a mismatch at the 3′ most subsequent position of a seed match. Thus, if the proportions of functional to non-functional target sites between the $S_{p,6}$ and the $S^{+}_{p,6}$ seed types were not significantly varying ($P>0.05$), the separation terminated otherwise the procedure was continued for the next seed type length. A P-value was calculated by means of a two-tailed Fisher's exact test. The α seed site separation terminated after three steps, the β seed matches contained two significant subsets and γ yielded no significant subsets. We termed the found significant seed types based on their start position and their length: $S_{p,k}=$'$k$mer$p$'. For standardization, we renamed the endmost subsets: $S^{+}_{\alpha,8}=8mer\alpha$, $S^{+}_{\beta,7}=7mer\beta$, $S^{+}_{\gamma,6}=6mer\gamma$.

To estimate the significance of our seed type set, we compared the distribution of the functional sites with a randomized pool of functional seed matches. We drew without replacement a subset of 7803 instances of the multinomial distribution Ω from functional and non-functional seed matches. A P-value was calculated by means of a χ^2 test of independence.

The impact of the seed types to miRNA target site prediction was evaluated in terms of recall and specificity. The recall estimates how many of the functional target sites OP are covered by a certain seed type $S_{p,k}$ and the specificity computes the fraction of correctly excluded non-functional target sites ON.

$$\text{Recall}(S_{p,k})=\frac{|\{s:s\in S_{p,k}\wedge s\in OP\}|}{|\{s:s\in OP\}|}$$

$$\text{Specificity}(S_{p,k})=\frac{|\{s:s\notin S_{p,k}\wedge s\in ON\}|}{|\{s:s\in ON\}|}$$

The recall and specificity of each miRNA target prediction algorithm was determined in terms of pure seed finding. Their seed type selection was assigned as described in the related literature. Due to ambiguous seed type assignments based on the first position of the target sequence, the specificity and recall values for TargetScan were computed by executing predictions on our mRNA set.

To estimate the miRNA seed type usage, we calculated the relative frequencies of a seed type for a certain miRNA. These values were normalized by the mean μ and the SD σ:

$$\text{Z-score}=\frac{x-\mu}{\sigma}$$

The conservation of each seed site was determined as described in (Betel *et al.*, 2008). We used the software package PHAST (Siepel *et al.*, 2005). The algorithm PhastCons is based on a phylogenetic hidden Markov model, which is fitted to the input sequence by maximum likelihood. Each nucleotide gets a score, which measures the evolutionary conservation across 17 vertebrates. For each seed match, the absolute chromosomal coordinates were determined and a conservation score was calculated. Only if the score of each nucleotide within a functional seed match exceeded the threshold of 0.57 (Betel *et al.*, 2008), the site was tagged conserved in mammals. The background conservation of a seed type was computed by calculating the fraction of conserved nucleotides of a non-redundant set of UTRs holding a specific seed type. For all statistical computations, the R programming language was applied (R Development Core Team, 2010).

3 RESULTS

3.1 Canonical seed types of miRNA target recognition

In this work, we defined a set of canonical seed types by analyzing the seed matches of experimentally verified functional target sites in the 3′UTR. The Ago HITS-CLIP miRNA:mRNA interaction map (murine assembly of 2006) (Chi *et al.*, 2009) lists 15 665 chromosomal positions of perfectly matching seed sites of length 6–8 nt belonging to 20 miRNAs frequently bound in Ago complexes. We mapped these sites to annotated protein-coding mRNA transcripts and retained sites located within the 3′UTR. For each miRNA, we scanned the 3′UTRs of the transcript set for all sites complementary to a miRNA subsequence beginning at either position one (α-position), two (β-position) or three (γ-position) relative to the miRNA 5′-end. We required a minimum length of 6 nt. Seed matches of length five [as reported by (Brennecke *et al.*, 2005)] were not significantly enriched in average Ago footprints (Supplementary Table S1). We classified these sites by means of their distance to an Ago HITS-CLIP site and retained miRNAs significantly enriched in footprints. This resulted in 2369 murine genes containing 7070 Ago HITS-CLIP sites of 16 broadly conserved miRNAs.

Each contiguous seed match was defined by its start position type and its length. The dataset was composed of eight α-, seven β- and five γ- seed match types (Supplementary Table S3). Following the law of Occam's razor, the simplest seed type setting for target prediction should usually be the correct one. To reduce unnecessary complexity of the seed type set, we identified unique seed types differing significantly from their superset in terms of functional and non-functional site distribution. For the murine and the human dataset, we achieved six different, disjunct types of seeds: three 6mers either beginning at the first nucleotide (6*mer*α), the second nucleotide (6*mer*β) or the third nucleotide (6*mer*γ), two 7mers

Table 1. Determined canonical seed types

	Functional		Non-functional			
Seed type	Frequency	%	Frequency	%	LOR[a]	P-value
6*merα*	1793	24	20 746	32	−0.12	$1.20E^{-028}$
6*merβ*	1382	19	13 500	21	−0.04	$2.57E^{-004}$
6*merγ*	1755	24	17 954	28	−0.06	$2.26E^{-009}$
7*merα*	760	10	5036	8	0.12	$2.03E^{-013}$
7*merβ*	959	13	5250	8	0.21	$1.34E^{-042}$
8*merα*	693	9	2203	3	0.44	$7.60E^{-132}$

[a]Log odds ratio based on sampling.

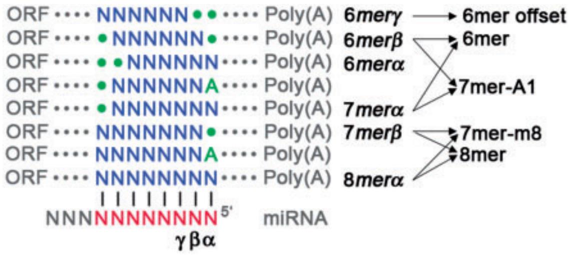
ORF •••• NNNNNN •• •••• Poly(A) 6mery → 6mer offset
ORF •••• • NNNNNN • •••• Poly(A) 6merβ → 6mer
ORF •••• • • NNNNNN •••• Poly(A) 6merα
ORF •••• • NNNNNNA •••• Poly(A) → 7mer-A1
ORF •••• • NNNNNNN •••• Poly(A) 7merα
ORF •••• NNNNNNN • •••• Poly(A) 7merβ → 7mer-m8
ORF •••• NNNNNNNA •••• Poly(A) → 8mer
ORF •••• NNNNNNNN •••• Poly(A) 8merα
NNNNNNNNNNN 5' miRNA
γβα

Fig. 1. Definition of seed types. The seed types were termed by the start position relative to the 5′-end of the miRNA and the length of the consecutive seed match. The defined set of canonical seed types can be surjectively projected to the seed type set of (Bartel, 2009). Equivalent definitions could be found for 6*merγ*, 6*merβ* and 7*merβ*. In the case of miRNAs having a seed sequence beginning with an uracile, 7*merα* complies with 7mer-A1 and 8*merα* is equal to 8mer. Otherwise 6*merβ* equates 7mer-A1 and 7*merβ* complies with 8mer. If the first position within the target sequence is not an adenine, 8*merα* equates 7mer-m8 and 7*merα* is equal to 6mer. Additionally, our set considered 6mer matches that are complementary to the first position of a miRNA seed (6*merα*). Common target site prediction tools focus on seeds of length seven and eight to increase precision.

either starting at position one (7*merα*) or position two (7*merβ*) and one 8mer beginning at the first nucleotide (8*merα*) (Supplementary Fig. S1A). These canonical seed types terminated within the first 8 nt of the miRNA in 97% of cases. This underscores the importance of the octamer at the miRNA 5′-end. The significance of this seed type set was evaluated by a sampling approach. The log odds ratio of long seed types is above zero, pointing to a better discrimination between functional and non-functional sites (Table 1).

In a previous work, (Bartel, 2009) defined seeds of miRNA target recognition. Comparing this previous definition with our canonical set of seed types, we recover this set and extend it by additional seed types starting at the α-position (Fig. 1). The miRNA seed match starting at position two and requiring a length of at least 6 nt was described as miRNA *core seed* (Bartel, 2009; Friedman *et al.*, 2009; Grimson *et al.*, 2007). In our canonical set, it is covered by seed types 6*merβ*, 7*merα*, 7*merβ* and 8*merα*.

3.2 Majority of functional sites are based on 6mer seeds

We examined recall and specificity affected by the individual seed types (Fig. 2). Focusing on the relative contribution of each seed type to functional sites, 6mer seeds make up the highest fraction of true target sites (recall: 0.67). On the other hand, 6mer types involve many false positives leading in sum to a low specificity (0.19) and precision (0.09, Supplementary Table S6). In terms of *in silico* target site classification, the usage of a short seed type causes an inverse prediction [Matthews correlation coefficient (MCC) < 0, Supplementary Table S6], suggesting the avoidance of such a type. In this case, reversing the classification would yield a result superior to an average random prediction.

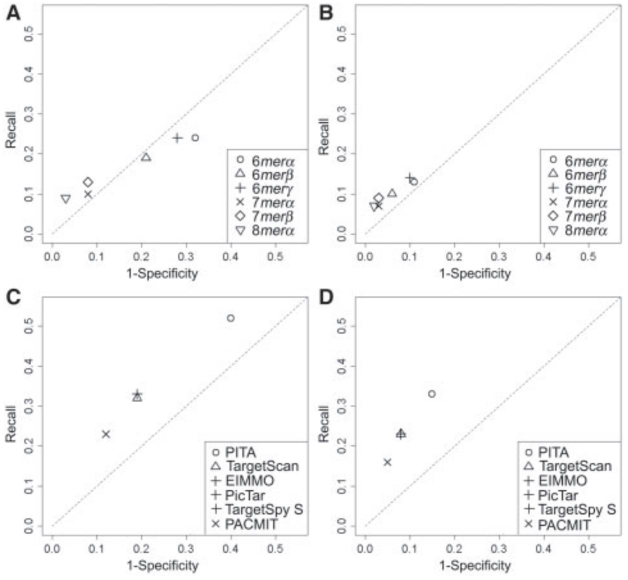

Fig. 2. Accuracy evaluation. (**A** and **C**) The impact of each seed type on miRNA target site prediction was determined by means of recall and specificity. The effect of the (default) seed type selection is shown for several prediction algorithms. These values present respectively the minimum specificity and the maximum recall of the tools. The dashed line shows an average random prediction. (**B** and **D**) Removing non-conserved target sites increases the precision, but lowers the recall. Note that panels (C and D) do not reflect the ranking of predictions based on the algorithms' scoring schemes.

Barely one-third of all genuine target sites are covered by seeds of length 7 and 8. Among these seed types, 7*merβ* holds the highest recall (0.13) and 8*merα* shows the best specificity (0.97). The combined set of 7- and 8mer matches achieves a specificity of 0.8 (precision: 0.19). miRNAs perform fine-tuning of gene expression, in particular 6mer seed matches are associated with low repressive effects (Friedman *et al.*, 2009). As most of the functional sites are formed by short seed sites, one can infer marginal reduction of the mRNA level to be the predominant effect of global miRNA-mediated regulation.

Computing recall and specificity in terms of miRNA:mRNA interactions resulted in a growth of both measurements for each seed type (Supplementary Fig. S2, Table S6 and S7). Here, only the presence of a site on a mRNA matters, whereas in terms of miRNA target site determination the location of a seed match relative to an Ago footprint is important. Multiple matches of one miRNA on a target mRNA are combined into one miRNA:mRNA interaction. Consequently, multiple false positive seed matches may be united to one true positive miRNA:3′UTR interaction. Conversely, multiple true negative target sites may be merged to one false positive interaction. Obviously, recall benefits but specificity suffers from these facts.

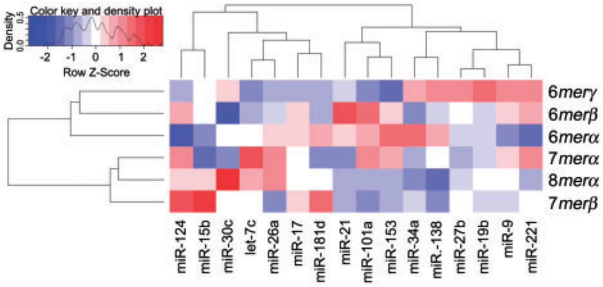

Fig. 3. Heatmap showing the seed type distribution for each miRNA. The colors affected by the row z-score indicate the bias of miRNAs to prefer targets holding a specific seed type. A red/blue coloration implies a higher/lower usage of a seed type compared to other miRNAs.

Moreover, we wondered if the seed type distributions differ between miRNAs. The relative frequencies of the seed types were computed for each miRNA. A z-score indicates miRNAs holding a frequency over or below the mean frequency given a specific seed type (Fig. 3). Interestingly, 6mer seed types and long seed types are grouped to clusters, respectively, demonstrating that a miRNA either binds to long sites or to short sites but not to both. Further, two main miRNA cluster appeared. The larger group contains miRNAs binding primarily to 6mer-based functional sites. Seven of the 16 miRNAs carry out stronger repression by pairing to rather long seed matches.

The importance of short seed types gains further support by the observation that 37% of the 3′UTRs contain exclusively seed matches of length six in their Ago footprints (Supplementary Table S4 lists the numbers of 3′UTRs containing seed matches of exclusively one type.) Interestingly, the sequences of this subset of 3′UTRs are significantly shorter than that of the superset (*t*-test, $P=4.53E^{-06}$). Stark *et al.* (2005) studied the impact of miRNA regulation on 3′UTR evolution and found that short 3′UTRs indicate avoidance of miRNA regulation. This goes well with our observation of short 3′UTRs regulated by less effective 6mer matches.

3.3 Non-conserved targeting relies on short seeds

We used a strategy established by (Betel *et al.*, 2008) to identify seed sites conserved across mammals (Fig. 4 and Supplementary Fig. S1B). The majority of functional target sites is conserved (60%). All seed types have a higher fraction of conserved sites than one would expect by chance, given the conservation of their 3′UTRs (Supplementary Table S5). The 6mer sites reveal an almost equal partitioning in conserved and non-conserved sites. A clear discrepancy between the numbers of conserved and non-conserved sites emerges for 7- and 8mer seeds. Particularly, 8*merα* seed matches exhibit a significant tendency to be conserved. The number of conserved sites in this case is more than three times as high as the number of non-conserved sites. In terms of 7mer seeds, about two-thirds of the seed matches are conserved, whereas 7*merα* exceeds 7*merβ*.

In summary, the mean probability to be conserved is about 55% for a 6mer seed. In contrast, 7mer and 8mer seeds have a probability up to 77% to be conserved. Further, a total of 75% of the functional non-conserved sites are covered by 6mer seeds. Therefore, non-conserved or species-specific targeting relies to a large extent on target sites containing short seeds.

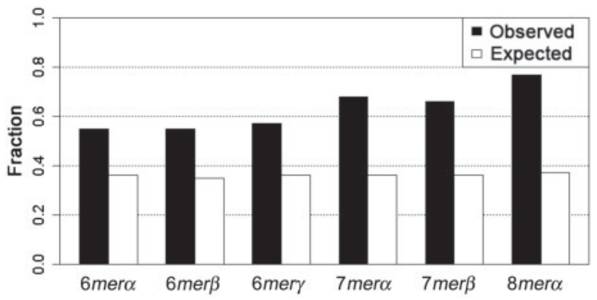

Fig. 4. Observed and expected fraction of conserved seed matches for each seed type illustrated for functional target sites.

Keeping only the conserved sites from the set of seed matches lifts specificity of all seed types (Fig. 2, Supplementary Table S9). In particular, the 6mer seeds show a significant increase of specificity leading to a classification better than an average random prediction (MCC > 0, Supplementary Table S6).

3.4 Target prediction focuses on 7- and 8mer seed matches

We reviewed frequently used approaches for target prediction in mammals with regard to the implemented seed types (Table 2). The TargetScan algorithm (Grimson *et al.*, 2007) seeks mainly for seeds of length seven and eight via seed types 7mer-A1, 7mer-m8 and 8mer. The 7mer-A1 sites may be of type 6*merβ* in the event the miRNA sequence starts with a nucleotide different to uracile. However, the majority of mammalian miRNAs begins with an U (Lewis *et al.*, 2005). Both PicTar (Krek *et al.*, 2005) and EIMMO (Gaidatzis *et al.*, 2007) require stringent seed pairing between 7 nt starting at either the *α* or the *β*-position. A novel approach called TargetSpy seed (Sturm *et al.*, 2010) restricts the set of seed matches to predictions containing a perfect 7mer.

Some algorithms allow for custom-defined seed searching: PITA (Kertesz *et al.*, 2007) seeks by default for sites of length six, seven and eight that start at position two of the miRNA. The standard setting of PACMIT (Marn and Vaníček, 2011) is even more restrictive by considering merely sites matching to miRNA positions two to eight. Both tools enable adjusting of the site length by the user. RNAhybrid (Krüger and Rehmsmeier, 2006) as well as IntaRNA (Busch *et al.*, 2008) are more flexible by providing a couple of additional parameters to customize the seed search, e.g. setting the start position. The latter is a general approach to predict RNA:RNA interactions. Both do not suggest default seed search parameters.

The impact of the (default) seed type selection of prediction algorithms on recall and specificity was evaluated (Fig. 2, Supplementary Tables S8 and S9).

Prediction methods implement scoring schemes to value target site characteristics beside the seed. In contrast to common evaluations of miRNA target site prediction algorithms, this is not an assessment of a subset of top scored instances but of all predictions. Therefore, the denoted specificity values represent the minima while the recall values show the maxima for the (default) seed choice, respectively. Subsets composed of top scored predictions would achieve significantly higher specificity values.

Obviously, all prediction models exhibit a considerable constraint regarding their ability of finding potential target sites. PITA holds

Table 2. Default miRNA seed type selection of prediction algorithms

	Seed type					
Algorithm	$6mer\alpha$	$6mer\beta$	$6mer\gamma$	$7mer\alpha$	$7mer\beta$	$8mer\alpha$
PITA[a]		✓		✓	✓	✓
TargetScan		✓[b]		✓[c]	✓	✓
PicTar				✓	✓	✓
EIMMO				✓	✓	✓
TargetSpy S.				✓	✓	✓
PACMIT[a]					✓	✓

[a] Configurable seed length, default seed types ensure high precision.
[b] If miRNA seed sequence starts with an adenine, guanine, cytosine.
[c] If miRNA seed sequence starts with an uracile.

the highest recall of 52% (specificity: 60%) owing to the exhaustive search for $6mer\beta$ seed matches, whereas PACMIT has the lowest recall of 23% (specificity: 88%) restricted to find less than a quarter of all functional seed sites. Additional filtering by removing conserved sites increases the specificity but consequently lowers the recall. Here, PACMIT could only find 16% of all functional sites (specificity: 73%). A higher recall but a lower specificity can be observed for the prediction of miRNA:mRNA interactions (Supplementary Fig. S2). Concluding, due to the significant gain of precision, tool developers recommend to use long seeds. Our study quantified the loss of recall accompanied by this proceeding.

4 CONCLUSION

In this study, we present an analysis of the most important feature for miRNA target recognition, the so-called miRNA seed, using a large-scale dataset of functional target sites. Based on the Ago HITS-CLIP and PAR-CLIP miRNA:mRNA interaction maps,we analyzed seeds properties and their influences on miRNA target site prediction methods. Due to the definite specification of Ago binding sites, we were able to classify miRNA recognition elements contained in the mRNA 3′UTR as either functional or non-functional. We defined a minimal set of seed types that is sufficient for accurate miRNA target site predictions. The final data pool allows for enhanced analysis of miRNA target prediction algorithms compared to earlier studies that were restricted by experimental constraints (Alexiou *et al.*, 2009; Selbach *et al.*, 2008). We found that most conserved miRNAs interact predominantly with target sites endowed with short seed matches; 67% of functional sites are based on 6mer seeds. In contrast, common prediction algorithms focus mainly on seeds of length seven or eight. At present, prediction algorithms have to accept severe deficiencies of recall to ensure high specificity that is naturally considered to be more important. Moreover, the preferential search for long seeds lifts the proportion of conserved sites. But we found that a substantial fraction (40%) of all functional target sites is not conserved. Target sites including 6mer seeds are enriched among these.

Concluding, the problem of recall can be easily translated to the problem of precision. However, this strongly intensifies the need for features beyond seed pairing that realistically describe miRNA targeting, in particular non-conserved target sites. It may also raise the basic question for the potential of seed-based approaches in discriminating between functional and non-functional sites.

Funding: Helmholtz Association within the Helmholtz Alliance on Systems Biology (project CoReNe); Federal Ministry of Education and Research (BMBF) within project SysMBo (Förderkennzeichen: 0315494A).

Conflict of Interest: none declared.

REFERENCES

Alexiou,P. *et al.* (2009) Lost in translation: an assessment and perspective for computational microrna target identification. *Bioinformatics*, **25**, 3049–3055.

Baek,D. *et al.* (2008) The impact of micrornas on protein output. *Nature*, **455**, 64–71.

Bagga,S. *et al.* (2005) Regulation by let-7 and lin-4 mirnas results in target mRNA degradation. *Cell*, **122**, 553–563.

Bartel,D.P. (2009) Micrornas: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Betel,D. *et al.* (2008) The microrna.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–D153.

Brennecke,J. *et al.* (2005) Principles of microrna-target recognition. *PLoS Biol.*, **3**, e85.

Busch,A. *et al.* (2008) Intarna: efficient prediction of bacterial srna targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.

Chi,S.W. *et al.* (2009) Argonaute hits-clip decodes microrna-mrna interaction maps. *Nature*, **460**, 479–486.

Farh,K.K.-H. *et al.* (2005) The widespread impact of mammalian micrornas on mRNA repression and evolution. *Science*, **310**, 1817–1821.

Friedman,R.C. *et al.* (2009) Most mammalian mrnas are conserved targets of micrornas. *Genome Res.*, **19**, 92–105.

Gaidatzis,D. *et al.* (2007) Inference of mirna targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, **8**, 69.

Griffiths-Jones,S. (2010) mirbase: microrna sequences and annotation. *Curr. Protoc. Bioinformatics*, **Chapter 12**, Unit 12.9.1–Unit 12.910.

Grimson,A. *et al.* (2007) Microrna targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Guo,H. *et al.* (2010) Mammalian micrornas predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.

Hafner,M. *et al.* (2010) Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, **141**, 129–141.

Hausser,J. *et al.* (2009) Relative contribution of sequence and structure features to the mRNA binding of argonaute/eif2c-mirna complexes and the degradation of mirna targets. *Genome Res.*, **19**, 2009–2020.

Karolchik,D. *et al.* (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.

Kertesz,M. *et al.* (2007) The role of site accessibility in microrna target recognition. *Nat. Genet.*, **39**, 1278–1284.

Krek,A. *et al.* (2005) Combinatorial microrna target predictions. *Nat. Genet.*, **37**, 495–500.

Krüger,J. and Rehmsmeier,M. (2006) Rnahybrid: microrna target prediction easy, fast and flexible. *Nucleic Acids Res.*, **34**, W451–W454.

Lewis,B.P. *et al.* (2003) Prediction of mammalian microrna targets. *Cell*, **115**, 787–798.

Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, **120**, 15–20.

Lim,L.P. *et al.* (2005) Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, **433**, 769–773.

Marn,R.M. and Vaníček,J. (2011) Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Res.*, **39**, 19–29.

Nielsen,C.B. *et al.* (2007) Determinants of targeting by endogenous and exogenous micrornas and sirnas. *RNA*, **13**, 1894–1910.

Pruitt,K.D. *et al.* (2009) NCBI reference sequences: current status, policy and new initiatives. *Nucleic Acids Res.*, **37**, D32–D36.

R Development Core Team (2010) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Selbach,M. *et al.* (2008) Widespread changes in protein synthesis induced by micrornas. *Nature*, **455**, 58–63.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Stark,A. *et al.* (2005) Animal micrornas confer robustness to gene expression and have a significant impact on 3′utr evolution. *Cell*, **123**, 1133–1146.

Sturm,M. *et al.* (2010) Targetspy: a supervised machine learning approach for microrna target prediction. *BMC Bioinformatics*, **11**, 292.

Supplementary methods

Preparation of PAR-CLIP

Hafner et al. (2010) identified clusters formed by at least 5 PAR-CLIP sequence reads and more than 20% T to C transitions. These 41 nt long regions were centered over the predominant crosslinking site. We mapped the chromosomal locations of 17,318 AGO1-4 crosslink centered regions (CCR) to the longest protein-coding mature mRNA transcript based on the NCBI reference sequence database annotation (Pruitt et al., 2009). All CCRs located within an exon of a mRNA 3'UTR (37 %) were retained. The dataset contained 580 miRNAs having at least one sequence read derived from the AGO PAR-CLIP. For miRNA families having the same seed sequence (position one to eight at the 5' end), we reduced the set to the member holding the highest sequence read count. Further, all non-conserved miRNAs (according to Friedman et al. (2009)) were removed. All mRNA and miRNA data was downloaded from UCSC (Karolchik et al., 2004) and miRBase (Griffiths-Jones, 2010) on January 2011.

We determined all sites complementary to a minimum of six contiguous nts beginning at either position one, two or three relative to the 5' end of the miRNA. Seed matches were classified functional or non-functional by means of their distance to the predominant crosslinking site. For each miRNA the seed match located within the CCR and nearest to its center was classified functional. Seed matches lying beyond the CCR were classified nonfunctional. To avoid false positives, additional miRNA target sites found within the CCR remained unclassified. Further, we retained only miRNAs whose target sites were significantly enriched in the CCRs. Transcripts having only non-functional sites were removed from the dataset. Finally, we got 21,214 functional, 380,893 non-functional and 665 unclassified seed sites for 72 miRNAs and 3,166 3' UTRs.

Formulae

For assessing the quality of the seed types and the miRNA target site prediction approaches we additionally used the following performance measures: precision and the Matthews correlation coefficient (MCC).

$$Precision = \frac{|TP|}{|PP|}$$

$$MCC = \frac{|TP| \cdot |TN| - |FP| \cdot |FN|}{\sqrt{|PP| \cdot |PN| \cdot |OP| \cdot |ON|}}$$

Given the set of observed functional (OP) and non-functional (ON) sites in the dataset and a set of predicted functional (PP) and non-predicted (PN) sites or a set of observed functional (OP) and non-functional (ON) miRNA seed type independent miRNA:mRNA interactions in the dataset and a set of predicted functional (PP) and non-predicted (PN) interactions we defined the confusion matrix as follows:

$$TP = PP \cap OP$$

$$FP = PP \cap ON$$

$$TN = PN \cap ON$$

$$FN = PN \cap OP$$

Supplementary tables

Table S 1. Enrichment of consecutive matching sites found in cluster peaks.

Site length	Sites in peak	Sites out of peak	Log odds ratio	P-value
5	14,876	208,562	0.00	$5.55E^{-001}$
6	6,239	54,346	0.21	$5.92E^{-289}$
7	2,295	14,772	0.34	$5.41E^{-281}$
8	948	3,963	0.53	$2.39E^{-279}$
9	219	983	0.50	$9.98E^{-059}$
10	45	242	0.42	$7.03E^{-010}$
11	13	58	0.50	$7.57E^{-005}$
12	2	16	0.25	$4.44E^{-001}$
13	2	3	0.97	$2.70E^{-003}$

Table S 2. Enrichment of miRNA seed matches in cluster peaks.

MiRNA	Sites in peak	Sites out of peak	Log odds ratio	P-value
miR-124	997	4102	0.53	$2.18E^{-301}$
miR-27b	932	6199	0.33	$4.71E^{-106}$
miR-9	752	4293	0.39	$1.48E^{-123}$
miR-181d	673	4904	0.29	$5.63E^{-060}$
miR-30c	647	5101	0.25	$3.30E^{-045}$
let-7c	643	2759	0.52	$3.80E^{-128}$
miR-15b	624	4690	0.27	$8.23E^{-051}$
miR-101a	607	4014	0.33	$8.78E^{-071}$
miR-26a	529	3831	0.29	$2.20E^{-048}$
miR-17	486	4677	0.17	$8.34E^{-016}$
miR-19b	478	4430	0.18	$1.85E^{-018}$
miR-34a	456	4278	0.18	$9.25E^{-017}$
miR-138	370	4087	0.11	$6.95E^{-006}$
miR-125b-5p	348	4425	0.04	$6.50E^{-002}$
miR-153	341	2753	0.24	$8.61E^{-023}$
miR-221	305	3073	0.15	$1.99E^{-008}$
miR-193	298	4123	0.01	$7.61E^{-001}$
miR-21	273	2543	0.18	$5.95E^{-011}$

Table S 3. All seed matching types found in the data sets.

Seed matching type			HITS-CLIP			PAR-CLIP		
Symbol	Start	Stop	Functional	Non-fun.	P-value	Functional	Non-fun.	P-value
$S_{\alpha,6}$	1	6	1,793	20,746	$2.20E^{-16}$	4,872	122,698	$2.20E^{-16}$
$S_{\alpha,7}$	1	7	760	5,036	$2.78E^{-09}$	2,254	31,090	$4.57E^{-07}$
$S_{\alpha,8}$	1	8	538	1,616	$4.08E^{-01}$	1,082	9,808	$8.15E^{-01}$
$S_{\alpha,9}$	1	9	115	429	$9.45E^{-01}$	264	2,424	$7.31E^{-01}$
$S_{\alpha,10}$	1	10	28	126	$6.84E^{-01}$	62	625	$9.30E^{-01}$
$S_{\alpha,11}$	1	11	10	25	$1.00E^{+00}$	18	184	$8.67E^{+01}$
$S_{\alpha,12}$	1	12	1	5	$1.00E^{+00}$	3	36	$6.96E^{+01}$
$S_{\alpha,13}$	1	13	1	2	$1.00E^{+00}$	0	14	$1.00E^{+00}$
$S_{\beta,6}$	2	7	1,382	13,500	$2.49E^{-08}$	4,583	83,728	$2.20E^{-16}$
$S_{\beta,7}$	2	8	720	3,998	$8.10E^{-01}$	2,321	22,924	$7.72E^{-02}$
$S_{\beta,8}$	2	9	181	935	$9.14E^{-01}$	547	6,631	$9.53E^{-01}$
$S_{\beta,9}$	2	10	47	247	$9.15E^{-01}$	136	1,582	$9.06E^{-01}$
$S_{\beta,10}$	2	11	9	49	$8.09E^{-01}$	33	416	$1.00E^{-00}$
$S_{\beta,11}$	2	12	1	14	$1.00E^{+00}$	9	105	$1.00E^{+00}$
$S_{\beta,12}$	2	13	1	7	$1.00E^{+00}$	1	21	$1.00E^{+00}$
$S_{\gamma,6}$	3	8	1,314	13,098	$5.03E^{-01}$	3,902	73,475	$6.54E^{-02}$
$S_{\gamma,7}$	3	9	339	3,647	$7.63E^{-01}$	851	18,769	$8.16E^{-01}$
$S_{\gamma,8}$	3	10	77	951	$8.15E^{-01}$	212	4,770	$8.15E^{-01}$
$S_{\gamma,9}$	3	11	20	199	$1.00E^{+00}$	48	1,183	$1.00E^{+00}$
$S_{\gamma,10}$	3	12	5	59	$1.00E^{+00}$	16	410	$1.00E^{+00}$

Table S 4. Distribution of UTRs holding canonical seed types of functional sites.

	Inclusive		Exclusive	
Seed type	Absolut	Relative	Absolut	Relative
$6mer\alpha$	1,179	0.50	195	0.08
$6mer\beta$	989	0.42	149	0.06
$6mer\gamma$	1,160	0.49	182	0.08
$7mer\alpha$	617	0.26	90	0.04
$7mer\beta$	769	0.32	122	0.05
$8mer\alpha$	577	0.24	101	0.04

Table S 5. Enrichment of functional sites in conserved regions (CR).

Seed type	Sites in CR	Sites out of CR	Log odds ratio	P-value
$6mer\alpha$	980	813	0.33	$1.23E^{-061}$
$6mer\beta$	755	627	0.34	$4.07E^{-050}$
$6mer\gamma$	1,000	755	0.36	$2.53E^{-071}$
$7mer\alpha$	515	245	0.57	$2.75E^{-072}$
$7mer\beta$	631	328	0.53	$3.33E^{-082}$
$8mer\alpha$	537	156	0.77	$1.12E^{-108}$

Table S 6. Overall quality assessment of seed types. The efficiency of determining a true miRNA target site and a true miRNA:mRNA interaction is considered.

	miRNA target sites		miRNA:mRNA interactions	
Seed type	Precision	MCC	Precision	MCC
$6mer\alpha$	0.08	-0.05	0.24	0.01
$6mer\beta$	0.09	-0.02	0.26	0.05
$6mer\gamma$	0.09	-0.03	0.25	0.04
$7mer\alpha$	0.13	0.03	0.29	0.06
$7mer\beta$	0.15	0.05	0.32	0.10
$8mer\alpha$	0.24	0.09	0.39	0.12

Table S 7. Overall quality assessment of seed types retaining conserved target sites.

	miRNA target sites		miRNA:mRNA interactions	
Seed type	Precision	MCC	Precision	MCC
$6mer\alpha$	0.12	0.02	0.30	0.08
$6mer\beta$	0.15	0.05	0.32	0.08
$6mer\gamma$	0.14	0.04	0.31	0.08
$7mer\alpha$	0.20	0.06	0.36	0.09
$7mer\beta$	0.23	0.08	0.39	0.11
$8mer\alpha$	0.32	0.11	0.46	0.13

Table S 8. Overall quality assessment of *in silico* approaches.

	miRNA target sites		miRNA:mRNA interactions	
Algorithm	Precision	MCC	Precision	MCC
PITA	0.13	0.07	0.28	0.12
TargetScan	0.16	0.10	0.31	0.15
PicTar	0.16	0.10	0.31	0.15
EIMMO	0.16	0.10	0.31	0.15
TargetSpy Seed	0.16	0.10	0.31	0.15
PACMIT	0.18	0.10	0.34	0.14

Table S 9. Overall quality assessment of *in silico* approaches retaining conserved target sites.

	miRNA target sites		miRNA:mRNA interactions	
Algorithm	Precision	MCC	Precision	MCC
PITA	0.20	0.15	0.35	0.18
TargetScan	0.24	0.15	0.38	0.18
PicTar	0.24	0.15	0.38	0.18
EIMMO	0.24	0.15	0.38	0.18
TargetSpy Seed	0.24	0.15	0.38	0.18
PACMIT	0.27	0.14	0.41	0.16

Supplementary algorithms

Algorithm S 1 FindCanonicalSeedTypes

```
Input: Start position type: $p$, seed match length: $k$, set of seed matches: $\Omega$, accumulator: $\Sigma$
Output: Significant seed types
//identify subsets
$S^{+}_{p,k} \leftarrow \{\forall s \in \Omega \ : \ starttype(s) = p \wedge length(s) = k\}$
$S_{p,k} \leftarrow \{\forall s \in \Omega \ : \ starttype(s) = p \wedge length(s) = k \wedge match(k+1) = false\}$
//test two-tailed statistical independence
if FishersExactTest($S^{+}_{p,k}$, $S_{p,k}$) $> 0.05$ then
    return $\Sigma$
else
    FindCanonicalSeedTypes($p$, $k+1$, $\Omega$, $\Sigma \cup S_{p,k}$)
end if
```

Supplementary figures

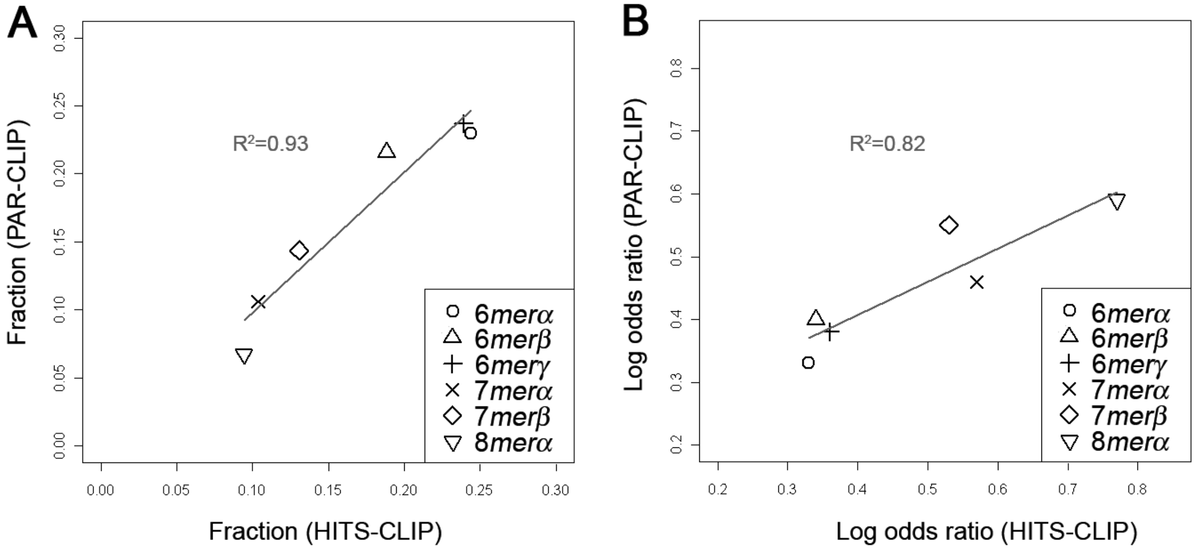

Figure S 1. Correlation of HITS-CLIP and PAR-CLIP. (A) The seed type distribution of functional sites is equal in both datasets (p-value: $2.1E^{-03}$). (B) To compare the conservation of functional sites the conservation of 3'UTRs was taken into account. The log odds ratio is equal in human and mouse (p-value: $1.3E^{-02}$).

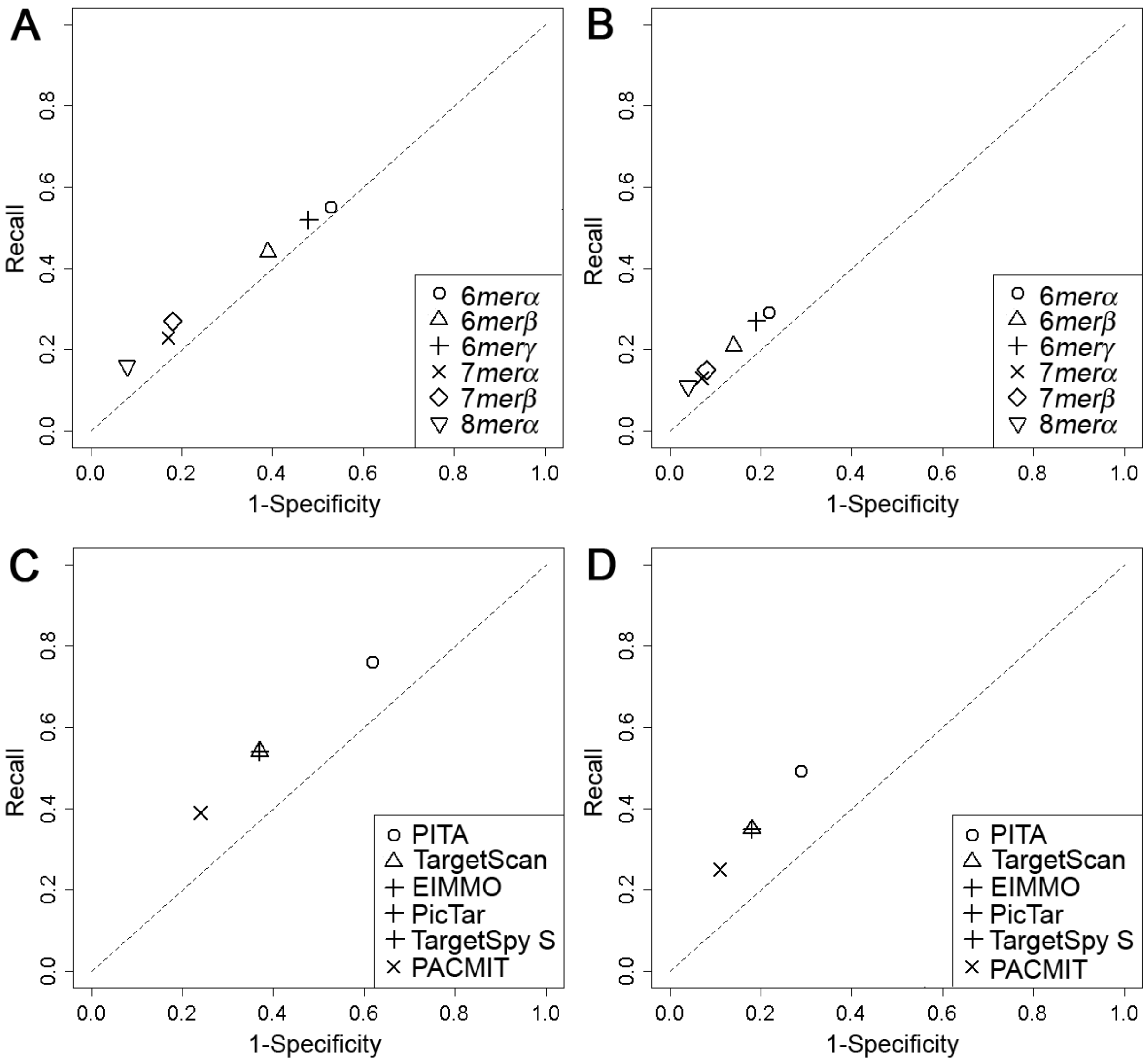

Figure S 2. Accuracy evaluation of miRNA:mRNA interaction determination. (A, C) The contribution of a seed type to miRNA:mRNA interaction prediction was measured by the receiver operating characteristic. The corresponding values for miRNA target prediction algorithms were determined. (B, D) The impact of retaining only conserved sites to each seed type/seed type subset of the prediction methods was computed.