Deposited research article
A statistical approach predicts human microRNA targets
Neil R Smalheiser and Vetle I Torvik

Addresses: University of Illinois at Chicago, UIC Psychiatric Institute, MC 912, 1601 W. Taylor Street, room 285 Chicago, IL 60612, USA

Correspondence: Neil R Smalheiser. E-mail: smalheiser@psych.uic.edu

A Statistical Approach Predicts Human microRNA Targets

Neil R. Smalheiser* and Vetle I. Torvik

University of Illinois at Chicago,

UIC Psychiatric Institute, MC 912,

1601 W. Taylor Street, room 285

Chicago, IL 60612

phone 312-413-4581

fax 312-413-4569

*corresponding author at: smalheiser@psych.uic.edu

Abstract

Background. MicroRNAs are ~18-24 nt. noncoding RNAs found in all eukaryotes that degrade messenger RNAs via RNA interference (if they bind in a perfect or near-perfect complementarity to the target mRNA), or arrest translation (if the binding is imperfect). Several microRNA targets have been identified in lower organisms, but no mammalian microRNA targets have yet been validated experimentally.

Results. We carried out a population-wide statistical analysis of how human microRNAs interact complementarily with human mRNAs present in the RefSeq database, looking for characteristics that differ significantly as compared with scrambled control sequences. These characteristics were used to predict a list of 72 candidate mRNA targets with 81% confidence. Unlike the case in *C. elegans* and *Drosophila*, many human microRNAs exhibited long exact matches (10 or more bases in a row), up to and including perfect target complementarity. Human microRNAs hit putative mRNA targets within the protein coding region about 2/3 of the time. And, microRNA hits in the candidate list did not have better complementarity near their 5'-end than expected by chance. In several cases, an individual microRNA hit multiple mRNAs that belonged to the same functional class.

Conclusions. The candidate list predicts a significant number of well-known and novel human genes that warrant experimental validation as mRNA targets, including several that may be regulated by RNA interference. The list also provides a training set and suggests an unified model to assist prediction of mRNA targets that do not have especially long regions of target complementarity.

Background

MicroRNAs (miRNAs) are small, ~18-24 nt. noncoding RNAs that are found in all
eukaryotes and are cleaved from larger ~70 nt. precursors via the action of Dicer enzyme
[reviews: ref. 1, 2]. MicroRNAs are thought to degrade messenger RNAs via eliciting
mRNA degradation (if they bind in a perfect or near-perfect complementarity to the
target mRNA), or to arrest translation of the mRNAs (if the binding complementarity is
imperfect). Although a number of microRNA targets have been identified in plants, *C.
elegans* and *Drosophila* [1, 2], few or no mammalian microRNA targets have yet been
validated, and there is reason to believe that the rules governing microRNA-target
interactions are not universal. For example, in plants, most of the known microRNAs
bind in a perfect or near-perfect manner to mRNA targets located within the protein
coding region (cds) [3, 4]. In contrast, in *C. elegans* [5] and *Drosophila* [6], known
microRNAs lack long stretches (>10) of complementarity with their targets and generally
interact within the 3'-untranslated region (3'-UTR). Furthermore, whereas the 5'-ends of
many *Drosophila* microRNAs recognize 5-6 nt. common motifs within the target, these
motifs are not a general feature of mammalian microRNAs [7].

In the present paper, we have performed a statistical analysis of the manner in which
human microRNAs interact complementarily with human mRNAs present in the NCBI
RefSeq database, looking for characteristics that differ significantly as compared with
scrambled versions of the same microRNA sequences. The results demonstrate several
novel features of human microRNA-mRNA interactions, and identify a short-list of

promising candidate microRNA-mRNA target pairs that are unlikely to have arisen by chance.

Results and Discussion

Population-wide statistical analyses were first carried out by examining the types of complementary interactions that occur between the set of microRNAs listed in Lagos-Quintana et al [8], and the set of human RefSeq mRNAs downloaded in August 2003. To obtain a list of individual candidate mRNA targets, analyses were repeated using all human microRNAs listed on the Sanger microRNA repository [9] and the set of human RefSeq mRNAs listed as of December 2003 [10]. To define the types of interactions that can occur by chance, ten independent sets of scrambled microRNA counterpart sequences were examined for complementarity with the mRNA population. Unless otherwise noted, the scrambled sequences were random permutations of the microRNA sequences, keeping constant the overall nucleotide composition. (Because microRNAs have a distinctive nonrandom di-nucleotide composition, we also confirmed that key findings were obtained when using scrambled sequences that had similar di-nucleotide composition to the microRNAs.)

1. Human microRNAs tend to have longer exact "hits" and fewer G:U matches than their scrambled counterparts.

First, we characterized the length distribution of exact complementarity between mRNA targets and nonredundant microRNAs (i.e. those that overlapped by 10 or more bases were collected into groups and the longest member of the group was chosen as nonredundant). MicroRNAs produced significantly longer exact "hits" on mRNAs than

their scrambled counterparts when G:U matches were excluded (fig. 1a). There was an excess number of hits in the microRNA set relative to scrambled control sequences at all exact hit lengths (10 or greater), becoming proportionately greater at longer hit lengths. When microRNAs were compared to scrambled sequences that matched the di-nucleotide composition of microRNAs, similar results were obtained. In contrast, this trend was not observed when G:U matches were included (not shown). Experimental studies suggest that RNA interference and arrested translation can still be elicited when small RNAs are modified to replace a number of Watson-Crick base pairs by G:U matches [11, 12]. On the other hand, G:U matches have distinctive binding energy and spatial orientation [13]. In the rest of the paper, "exact hits" will refer to complementarity without G:U matches.

We then examined the extended target interactions produced when microRNA sequences were first lined up with targets according to each exact hit ($\geq$ 10 bases), and then allowed to extend in both directions along the target, either with or without permitting G:U matches. A modified gapped-BLAST algorithm [14] was used to compute the optimal alignment, employing a weighted score that takes gaps and mismatches into account (r=10, q=-2.5, G=8, E=0.5). Without permitting G:U matches in the extension phase, microRNAs had better average gapped-BLAST scores than scrambled counterparts (153.00 ± 0.03 vs. 150.98 ± 0.01, mean $\pm$ s.e.m., p<0.0001). With permitting G:U matches in the extension phase, the microRNA set showed significantly fewer G:U matches overall relative to scrambled counterparts, even when holding constant the length of the exact hit (2.891 ± 0.004 vs. 2.939 ± 0.001, p<0.0001). Taken together, these findings indicate that the population of microRNAs exhibits better

complementarity to mRNAs than their scrambled counterparts, and tend to use fewer G:U

basepairs than expected by chance.

In lower organisms, individual validated microRNA targets tend to receive multiple

hits by distinct microRNAs [1, 2]. In humans too, individual mRNA targets were hit by

multiple nonredundant microRNAs more often than by their scrambled counterparts, and

this was particularly striking when the hits were located close together (fig. 1b).

2. Identifying the candidate mRNA target list.

When combined, exact hit length, gapped-BLAST score and presence of multiple

distinct hits on the same mRNA target gave better discrimination power than any single

feature, suggesting that they give insight into biologically true mRNA targets.   We

examined four different combinations of these parameters, each chosen to maximize the

total number of candidates while keeping the discrimination ratio as high as possible.

Candidate list #1 was generated simply by defining a cut-off of 17 exact hit length;

this alone could discriminate well between targets hit by the microRNA set vs. the

scrambled sets (fig. 1a; 14 vs. an average of 1.9, or a ratio of 7.4 to 1). A similar

discrimination ratio was observed when comparing scrambled sequences maintaining the

same di-nucleotide composition as the microRNAs. List #2 consisted of targets with

multiple hits from distinct microRNAs less than 25 bases apart, with at least one exact hit

$\geq$13 bases and with at least one gapped BLAST score $\geq$ 185 (not counting G:U), <u>or</u> that

exhibited a perfect complementarity including G:U matches. For the next two lists, we scored only exact hits $\geq$10 bases long and that occurred $\leq$ 50 times within the entire mRNA population; this minimized "noise" arising from common or low-complexity target sequences, albeit at the cost of removing some target sequences that are shared within protein families. List #3 required two or more hits from distinct microRNAs $\leq$ 100 bases apart, at least one exact hit $\geq$14 bases and one gapped-BLAST score of $\geq$ 190 (not counting G:U). List #4 required hits $\leq$ 500 bases apart, at least one exact hit $\geq$ 14 bases, and at least one gapped-BLAST score $>$ 89% of the best possible score including G:U matches (this takes into account the fact that longer microRNAs have greater possible absolute scores than shorter microRNAs).

Because all four candidate lists had some overlapping members, and had similar characteristics, they were combined into a single list consisting of 72 candidate mRNA targets (Table 1), hit by almost the entire set of nonredundant microRNAs (i.e., 107/109). In contrast, scrambled counterpart sequences hit an average of 13.7 $\pm$ 1.15 targets and were represented by 54.3 $\pm$ 3.5 nonredundant sequences. The candidate list gives an overall discrimination ratio of 5.3 to 1, meaning that 81% should be accurately assigned as true targets for one or more microRNAs. See additional data file 1 for a fully annotated candidate mRNA target list, additional data file 2 for all microRNA hits upon the candidate list (extended with and without including G:U matches), and additional data file 3 for a list of the nonredundant microRNAs together with their putative targets.

3. Characterizing the candidate list.

The mRNA targets on the candidate list had a larger number of hits per kilobase of target sequence than did the scrambled counterparts (2.17 ± 0.1 vs. 1.83 ± 0.085, p=0.006). As well, individual microRNAs hit multiple (up to 17) distinct members of the candidate list, which again happened significantly more often than by chance (fig. 2). Surprisingly, there was no preference for microRNA hits to be located within 3'-untranslated regions: 5% of hits were located in the 5'-UTR, 1% at the 5'-UTR/coding junction, 67% in the protein coding region, 1% at the coding/3'-UTR junction, and only 26% in the 3'-UTR. This distribution was not significantly different from hits produced by the scrambled sequences. Finally, microRNAs did not have relatively better target complementarity near their 5'-end: Only 13% of hits had ≥ 7 exact hit length starting at position 1 or 2 relative to the 5'end of the microRNA (vs. 17.5% of hits produced by scrambled sequences).

3. Specific mRNA candidates with high face validity include those with perfect and near-perfect complementarity.

1. miR-196 hit the mammalian homeobox gene HOXB8 with perfect complementarity (22/22) including G:U matches (Table 2). This hit occurred in the 3'-UTR, in a region of open secondary structure upon the mRNA, and is identical between man and mouse (not shown). Thus, this is an ideal candidate by anyone's criteria. HOXB8 and EphA5 were both hit by the same microRNA, miR-198 -- this is interesting since both genes are involved in e.g., hindbrain patterning, and since mammalian homeobox genes are known to regulate Eph and ephrin expression [15].

2. Five additional members of the candidate list received microRNA hits with either perfect complementarity or one mismatch (Table 2). One of these is a retrotransposon-encoded reverse transcriptase that has been previously reported to have 2 microRNA precursors encoded on its opposite strand, that may regulate its expression [16, 17]. Although the others are mRNAs of unknown function that showed no sequence homology amongst themselves, the group of perfect and near-perfect mRNA targets, together with two other mRNAs, exhibited a strong apparent co-regulation by microRNAs (Table 3). MiR-133a hit 5 members of this group (out of a total of 6 hits in the candidate list), and 11 other microRNAs hit 2 or more members. In fact, two pairs of mRNAs shared 3 microRNAs (Table 3). This strongly suggests that this set of mRNAs is functionally related in some manner.

3. Four additional mRNAs on the candidate list are related to reverse transcriptase, and all were hit by the same set of 4 microRNAs (29b, 136, 145 and 223). This is consistent with evidence that a major role for RNA interference is thought to be to counteract transposon mobilization [review: 18].

4. MicroRNA 145 hits 17 targets on the candidate list, of which a disproportionate number (6) are in the signal transduction category and three of these are related to GTPase activation (Rho GTPase-activating protein (RICS), G protein gamma 7, and hypothetical protein FLJ32810 – containing RhoGAP and SH3 domains; Table 1). A recent study showing that miR-143 and miR-145 are both underexpressed in colorectal

neoplasia [19] had previously proposed the first two of these candidates as potential

targets. Interestingly, the third target found here is not only novel (XM_350859,

RhoGAP-like) but is hit by both miR-143 and miR-145 in close proximity (see additional

data file 2), further suggesting that this is likely to be a true biological target for

microRNA regulation.

Conclusions

A combination of three simple parameters was sufficient to create a list of 72 human

mRNA candidates, such that 81% are expected to represent true microRNA targets

(Table 1). By comparing how the population of microRNAs vs. their scrambled

counterparts interact with the population of human RefSeq mRNA sequences, we

estimate that the probability of detecting a true microRNA target increases a) as the

length of exact complementarity of a "hit" between microRNA and target increases, b) as

the overall complementarity of a "hit" increases (allowing for gaps, G:U matches and

mismatches), and c) as two distinct microRNAs hit the same mRNA in closer proximity.

Targets on the candidate list also received more hits per unit length and more multiple

hits from distinct microRNAs than expected by chance.

While this manuscript was being written up, four different papers appeared that used

computational approaches to predict microRNA targets in *Drosophila* [20-22], and

mammals [23], using different strategies, criteria and filters than employed here. In

particular, these studies only considered hits occurring within 3'-UTR regions that were

conserved across related species, and favored or required a short region of perfect

complementarity towards the 5'-end of microRNAs. As a consequence, there is little

overlap between specific entries on our candidate list and those reported by others. Nevertheless, very similar types of targets were predicted across studies, including members of the same gene families. Transcription factors (including homeobox genes) and nucleic acid-binding proteins are among the top predicted microRNA targets. As well, many other functional categories are represented including kinases, receptors and other signal transduction proteins, membrane and cytokeletal proteins, and effectors of differentiation (Table 1).

The candidate list is ready for immediate experimental verification in tissues where both the microRNA and the putative target are co-expressed, and we hope that investigators will find it worthwhile to examine their favorite genes. (Note, however, that RNA A-to-I editing (24, 25) might prevent some potential targets from being operative in vivo.) It is possible that additional predictive rules may yet be uncovered for the majority of mRNA targets that do not have long exact hit lengths (e.g. specific placement of G:U matches or gaps, participation of RNA-binding proteins, and secondary structure of the mRNA target region [12]; see also [23]). It is also likely that additional microRNA targets exist within transcribed sequences not included in RefSeq, and possibly within non-transcribed genomic sequences as well.

The candidate list represents the few best predicted targets, yet the population-wide characteristics of human microRNA-mRNA interactions support and extend the notion that microRNAs are likely to form extensive gene-regulatory networks [26]. If microRNAs simply bind more or less well to many potential mRNA targets, then single

interactions with relatively better binding affinity will persist longer on the target and hence have greater regulatory impact. Multiple non-overlapping hits from different microRNAs should have an avidity effect, increasing the overall probability that at least one hit region is occupied upon the target. Such a model implies that increasing the abundance of a particular microRNA in a cell would expand the set of its mRNA targets by recruiting those that are marginally effective.

A back-of-the-envelope calculation suggests that each microRNA may have at least ~10 mRNA targets in the human transcriptosome. (The excess number of exact hits ≥ 10 bases long in RefSeq (microRNA set minus the scrambled set) is ~25,000 (fig. 1a); this is divided by ~8 hits per mRNA observed in the candidate list and ~250 microRNAs. Not all such hits will be onto true targets, but this is offset by the presence of many true hits having less than 10 bases in a row [23].) Thus, even if microRNA sequences are generally tightly constrained during the course of evolution because they hit a few key conserved mRNA targets, this does not imply that most of their mRNA targets are similarly constrained. The extensive redundancy among microRNAs presumably is a way to regulate interactions with different targets in a tissue- and developmental stage-specific manner. This view suggests that as new species-specific mRNAs arise during the course of evolution, new species-specific targets (and new regulatory functions) are expected to be recruited frequently.

Abbreviations

5'-UTR, 5'-untranslated region. CDS, protein coding region. 3'-UTR, 3'-untranslated region.

Methods

MicroRNAs. Statistical analyses were first carried out using the set of mouse and human microRNAs listed in Lagos-Quintana et al [8], and then repeated to obtain individual candidate mRNA targets using all human microRNAs listed on the Sanger microRNA repository [9] as of December 2003. These sources were combined to create nonredundant microRNA sets (i.e. microRNAs that have 10 or more consecutive nucleotides in common were collected into groups and the longest member of the group was chosen as nonredundant). Additional data file 4 lists the nonredundant microRNAs, together with the corresponding redundant microRNAs in each group. Almost all mouse microRNAs have exact human counterparts, but hits were annotated with mouse entries in cases of minor corrections and discrepancies between these two sources. One individual microRNA (mir-207) and several scrambled sequences were found to be low-complexity or complementary to abundant repeats (e.g., Alu) and were removed from consideration.

mRNAs. Analyses were first carried out using the set of human RefSeq mRNAs available in August 2003, and then supplemented with additional human RefSeq mRNAs listed as of December 2003. A) Sequences in RefSeq > 20,000 bases long were removed from consideration because they were hit by many, if not all microRNAs, and a few sequences > 15,000 bases long were removed from the final candidate list because they had a relatively high false-positive probability. B) When counting the number of hits over the population of mRNAs, two hits were counted as redundant if the entire region around the hit (plus or minus 25 nucleotides on each side) was identical. C) When counting

distinct hits by microRNAs on the same target, two hits were counted as redundant if they

shared the same exact hit. This minimized possible artifacts due to overlapping

microRNAs, as well as removed cases in which microRNAs hit exactly-repeating

sequences within the target. D) In tabulating hits onto mRNA targets, we did not count

hits that contained low-complexity sequences as detected by the DUST algorithm

encoded by a Perl script provided by Lincoln Stein [27]. E) When assembling the

candidate mRNA target list, we chose a single exemplary mRNA and removed other

entries that were transcript variants or nearly identical by BLAST searching. In the

course of this study, some of the target mRNAs were removed from RefSeq for routine

genome annotation processing. If these were subsequently replaced with updated

versions of these mRNAs in RefSeq that included the same hits, the latter version is listed

here as well. For those entries removed but not replaced in RefSeq at the time of

submission of the manuscript, other active entries currently in Genbank are listed if

possible.

Statistics. To decide whether the number of observed microRNA hits were

significantly different from chance, 10 replications of scrambled sequences were used to

estimate prediction intervals. The prediction interval allows one to say with 95%

confidence that any single new replication of the scrambled set will be below the value of

the microRNA set. Prediction intervals were chosen as more conservative and more

appropriate than confidence intervals.

Acknowledgements

References

1. Lai EC: **microRNAs: runts of the genome assert themselves.** *Curr Biol* 2003, **13**R925 -936.

2. Carrington JC, Ambros V: **Role of microRNAs in plant and animal development.** *Science* 2003, **301:**336-338.

3. Llave C, Xie Z, Kasschau KD, Carrington JC: **Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA.** *Science* 2002, **297:**2053-2056.

4. Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, Bartel DP: **Prediction of plant microRNA targets.** *Cell* 2002, **110:**513-520.

5. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of Caenorhabditis elegans.** *Genes Dev* 2003, **17:**991- 1008.

6. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, Gaasterland T, Meyer J, Tuschl T: **The small RNA profile during Drosophila melanogaster development.** *Dev Cell* 2003, **5:**337-350.

7. Lai EC: **Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation.** *Nat Genet* 2002, **30:**363-364.

8. Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T: **Identification of tissue-specific microRNAs from mouse.** *Curr Biol* 2002, **12:**735 -739.

9. **The miRNA Registry**

[http://www.sanger.ac.uk/Software/Rfam/mirna/]

10. **RefSeq**

[http://www.ncbi.nlm.nih.gov/RefSeq/]

11. Pusch O, Boden D, Silbermann R, Lee F, Tucker L, Ramratnam B: **Nucleotide sequence homology requirements of HIV-1-specific short hairpin RNA.** *Nucleic Acids Res* 2003, **31:**6444 -6449.

12. Kretschmer-Kazemi Far R, Sczakiel G: **The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides.** *Nucleic Acids Res* 2003, **31:**4417-4424.

13. Varani G, McClain WH: **The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems.** *EMBO Rep* 2000, **1:**18-23.

14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389- 3402.

15. Wilkinson DG: **Multiple roles of EPH receptors and ephrins in neural development.** *Nat Rev Neurosci* 2001, **2:**155-164.

16. Seitz H, Youngson N, Lin SP, Dalbert S, Paulsen M, Bachellerie JP, Ferguson-Smith AC, Cavaille J: **Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene.** *Nat Genet* 2003, **34:**261-262.

17. Smalheiser NR: **EST analyses predict the existence of a population of chimeric microRNA precursor-mRNA transcripts expressed in normal human and mouse tissues.** *Genome Biol* 2003, **4:**403.

18. Zamore, PD: **Ancient pathways programmed by small RNAs.** *Science* 2002, **296:**1265-1269.

19. Michael MZ, O' Connor SM, van Holst Pellekaan NG, Young GP, James RJ: **Reduced accumulation of specific microRNAs in colorectal neoplasia.** *Mol Cancer Res* 2003, **1:**882-891.

20. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: **MicroRNA targets in** *Drosophila. Genome Biol* 2003, **5:**R1.

21. Stark A, Brennecke J, Russell RB, Cohen SM**: Identification of Drosophila MicroRNA Targets.** *PLOS Biol* 2003, **1:**397-409.

22. Rajewsky N, Socci ND: **Computational identification of microRNA targets.** *Dev Biol*, in press.

23. Lewis BP, Shih I-h, Jones-Rhoades MW, Bartel DP, Burge CB: **Prediction of Mammalian MicroRNA Targets.** *Cell* 2003, **115:**787-798.

24. Scadden AD, Smith CW: **RNAi is antagonized by A-->I hyper-editing.** *EMBO Rep* 2001, **2:**1107-1111.

25. Tonkin LA, Bass BL: **Mutations in RNAi rescue aberrant chemotaxis of ADAR mutants.** *Science* 2003, **302:**1725.

26. Ke XS, Liu CM, Liu DP, Liang CC: **MicroRNAs: key participants in gene regulatory networks.** *Curr Opin Chem Biol* 2003, **7:**516-523.

27. **Bioperl: Repetitive DNA**
[http://bioperl.org/pipermail/bioperl-l/1999-November/003313.html]

Figure 1. microRNAs and their scrambled counterparts interact differently with the population of human mRNAs. Shown are all exact hits $\geq$ 10 bases long (not counting G:U matches) produced on human RefSeq mRNAs by the set of nonredundant microRNAs, vs. the average of 10 replications of scrambled control sequences. A) Number of hits as a function of exact hit length. Only the longest hit was counted: e.g., for a hit of length 18, the two subsets of length 17 in the same hit position were not counted. B) Number of distinct mRNA sequences which received hits from two or more distinct microRNAs, as a function of the minimum distance between hits. (Distance of 0 or 1 was excluded because this might be produced by partial overlap of microRNA sequences.)

Figure 2. Individual microRNAs hit multiple targets on the candidate list, more often than expected by chance.

Table 1 – The Candidate mRNA Target List

Transcription factors and other nucleic-acid binding proteins (15)

homeo box B8 (HOXB8)
E2F transcription factor 6 (E2F6)
transcription factor 20 (AR1) (TCF20)
DEAD (Asp-Glu Ala-Asp) box polypeptide 51 (DDX51)
similar to ATP-dependent RNA helicase DDX24 (DEAD-box protein 24) (LOC221311)
myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog); translocated to, 1
    (MLLT1)
high mobility group AT-hook 2 (HMGA2)
polymerase (DNA directed), theta (POLQ)
strand-exchange protein 1 (SEP1)
hypothetical protein FLJ12994 - RFX DNA-binding domain
similar to LINE-1 reverse transcriptase homolog (LOC285907)
similar to hypothetical protein (L1H 3 region) – related to reverse transcriptase
similar to putative p150 (LOC282945) – related to reverse transcriptase
similar to reverse transcriptase related protein (LOC222252)
similar to RTl1 (LOC376283) – related to reverse transcriptase

Kinases, receptors and other signaling proteins (13)

fyn-related kinase (FRK)
WNK kinase, lysine deficient 3 (PRKWNK3)
protein phosphatase 2, regulatory subunit B (B56), epsilon isoform (PPP2R5E)
EphA5 receptor (EPHA5)
killer cell lectin-like receptor subfamily A, member 1 (KLRA1)
polycystin and REJ (sperm receptor for egg jelly homolog, sea urchin)-like (PKDREJ)
integrin, alpha X (antigen CD11C (p150), alpha polypeptide) (ITGAX)
inositol 1,4,5-triphosphate receptor, type 1 (ITPR1)
hypothetical protein FLJ32810 - RhoGAP domain, SH3 domain
hypothetical protein FLJ00058 – G protein gamma 7
Rho GTPase-activating protein (RICS)
hypothetical protein FLJ30899 – probable ras GAP
similar to ADP-ribosylation factor-like membrane-associated protein (LOC132946) -
    ARF-like small GTPase domain, Sar1p-like member of the Ras-family

Membrane and extracellular proteins (11)

Laminin, beta 4 (LAMB4)
laminin, gamma 2 (LAMC2)
fibronectin 1 (FN1)
collagen, type IV, alpha 5 (Alport syndrome) (COL4A5)
collagen, type XIX, alpha 1 (COL19A1)
similar to Voltage-dependent anion-selective channel protein 1 (VDAC-1)

ATPase, Na+/K+ transporting, alpha 2 (+) polypeptide (ATP1A2)
complement component 1, q subcomponent, beta polypeptide (C1QB)
hypothetical protein FLJ20506 – transmembrane protein
MAM domain containing glycosylphosphatidylinositol anchor 1 (MDGA1) – Ig, MAM
    domains
similar to TCAM-1 (LOC284171)

Cytoskeletal domain-containing proteins (7)

myosin heavy chain Myr 8 (MYR8)
ankyrin repeat domain 17 (ANKRD17)
KIAA1817 protein – intermediate filament, ATPase, PDZ, Band 4.1, FERM domains
chromosome 10 open reading frame 39 (C10orf39) – homologous to myosin, plectin
oxysterol binding protein 2 (OSBP2) – pleckstrin homology domain
KIAA1202 protein – PDZ, ATPase domains
hypothetical protein FLJ23529 - homolgous to dynein heavy chain

Miscellaneous or unknown function (26)

cell cycle progression 2 protein (CPR2)
olfactomedin 3 (OLFM3)
histidine rich calcium binding protein (HRC)
interferon-related developmental regulator 1 (IFRD1)
KIAA1301 protein - NEDD4-related E3 ubiquitin ligase NEDL2
KIAA1203 protein - ubiquitin C-terminal hydrolase
hydroxyprostaglandin dehydrogenase 15-(NAD) (HPGD)
UDP-GlcNAc:betaGal beta-1,3-N-acetylglucosaminyltransferase 1 (B3GNT1)
KIAA1854 protein - leucine rich repeat C-terminal domains
testis specific, 14 (TSGA14)
chromosome 4 open reading frame 1 (C4orf1) –membrane AND nuclear protein
hypothetical protein FLJ33069
hypothetical protein FLJ32731
hypothetical protein FLJ38464
hypothetical protein LOC285431
hypothetical protein LOC284107
similar to agCP1362 [Anopheles gambiae str. PEST] (LOC344751)
KIAA1632 protein
similar to hypothetical protein D11Ertd497e (LOC343360)
LOC138724
LOC343460
LOC340963
LOC343220
LOC285842
LOC352767
LOC350293

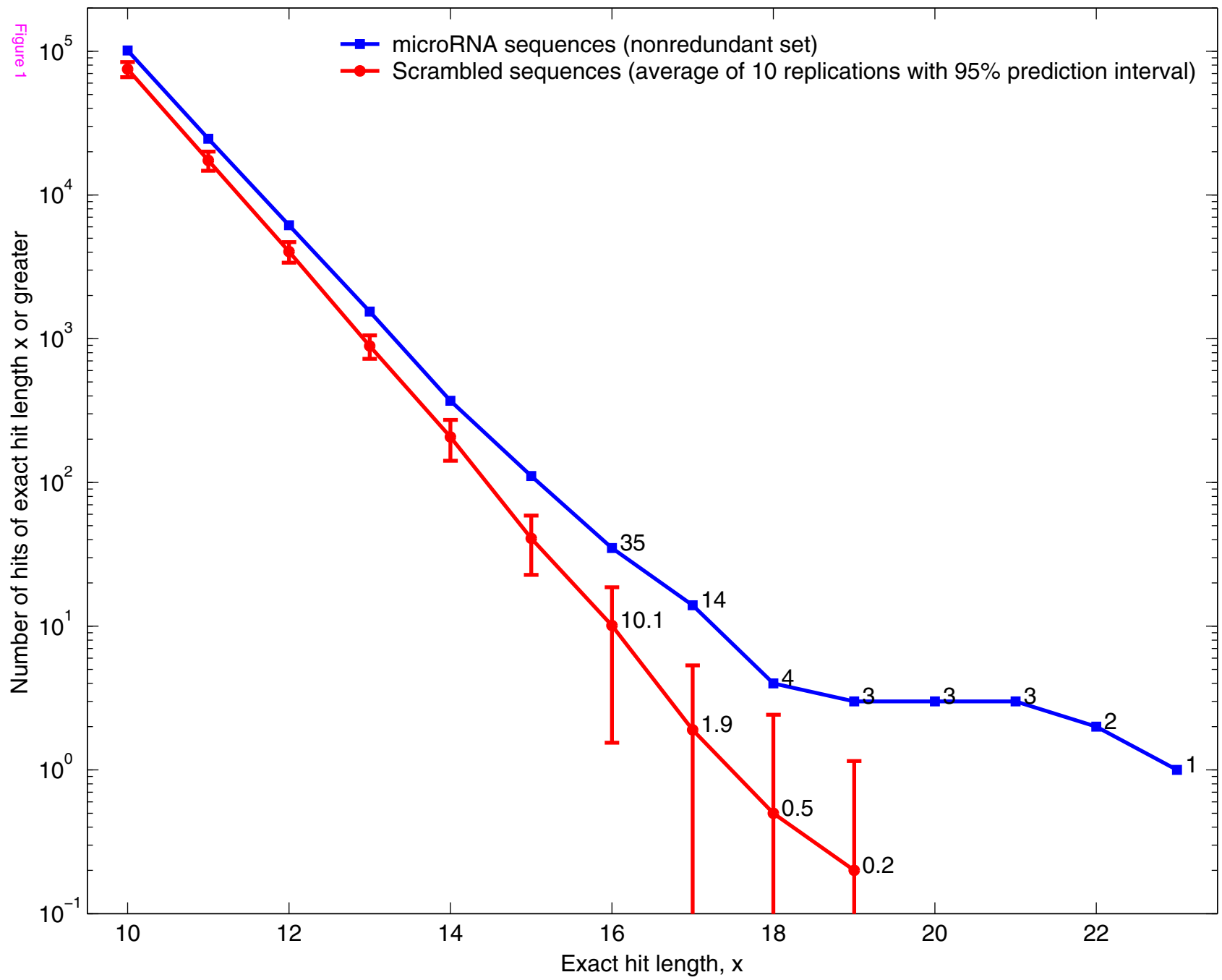**Table 2. Cases of perfect or near-perfect complementarity between microRNAs and their candidate targets.**

```
# 1) XM_352144.1 - similar to RTl1 (LOC376283)
hsa-mir-136   3' AGGUAGUAGUUUUGUUUACCUCA  5'
                 |||||||||||||||||||||||
RTl1          5' UCCAUCAUCAAAACAAAUGGAGU  3'
                 51        CDS        73/4077


hsa-mir-127   3' UCGGUUCGAGUCUGCCUAGGCU 5'
                 |||||||||||||||||||||
RTl1          5' AGCCAAGCUCAGACGGAUCCGA 3'
                 1733       CDS       1754/4077



# 2) XM_305931.1 - LOC352767
hsa-miR-219   3' UCUUAACGCAAACCUGUUAGU 5'
                 ||||||||||||||||||||
LOC352767     5' AGAAUUGCGUUUGGACAAUCA 3'
                 109       CDS        129/915



# 3) NM_024016.2 - homeo box B8 (HOXB8)
hsa-mir-196   3' GGUUGUUGUACUUUGAUGGAU 5'
                 |||||||||||||||||•||||
HOXB8         5' CCAACAACAUGAAACUGCCUA 3'
                 1379      3'UTR     1399/1823



# 4) XM_351779.1 - hypothetical protein FLJ32731
hsa-mir-185   3' CUUGACGGAAAGAGAGGU 5'
                 ||•|||||||||||||•|•
FLJ32731      5' GAGCUGCCUUUCUCUUCG 3'
                 1128      CDS   1145/4156



# 5) XM_211898.1 - hypothetical protein LOC285431
hsa-mir-95    3' ACGAGUUAUUUAUGGGCAACUU 5'
                 |||||||||||||••••||||||
LOC285431     5' UGCUCAAUAAAUGUUUGUUGAA 3'
                 2407      3'UTR    2428/2466



# 6) XM_303960.1 - LOC350293
hsa-mir-146   3' UUG-GGUACCUUAAGUCAAGAGU 5'
                 |•• ||||||||||||||||••
LOC350293     5' AGUACCAUGGAAUUCAGUUCUUG 3'
                 324        CDS        346/1066



# 7) NM_144649.1 - hypothetical protein FLJ33069
hsa-mir-133a  3' UGUCGACCAACUUCCCCUGGUU 5'
                 |||  |||||||||||||||•
FLJ33069      5' ACAACUGGUUGAAGGGGACCAG
                 520        CDS        541/1993
```

Table 3 – A cluster of genes with an unusually large number

of microRNAs in common.

mRNA \ miR	133a	149	182	136	122b	186	26b	9*	130a	214	196	198
FLJ33069	1	1										
RTl1	1	2	1	1	1							
LOC343460	1		1	1		1	1					
LOC285431	1					1		1	2			
KIAA1632	1						1	2	1	2	1	
FLJ32731					1		1			2		1
HOXB8			1								1	1

Numbers indicate how many times each microRNA hits each mRNA target.

Figure 1

Figure 2

Figure 3