Genomics & Informatics Vol. 3(3) 53-62, September 2005

Prediction of Mammalian MicroRNA Targets - Comparative Genomics Approach with Longer 3' UTR Databases

Seungyoon Nam[1,2], Young-Kook Kim[3], Pora Kim[1], V. Narry Kim[3], Seokmin Shin[2], and Sanghyuk Lee[1],*

[1]Division of Molecular Life Sciences, Ewha Womans University, Seoul 120-750, Korea
[2]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Korea
[3]Institute of Molecular Biology and Genetics and School of Biological Science, Seoul National University, Seoul 151-742, Korea

Abstract

MicroRNAs play an important role in regulating gene expression, but their target identification is a difficult task due to their short length and imperfect complementarity. Burge and coworkers developed a program called TargetScan that allowed imperfect complementarity and established a procedure favoring targets with multiple binding sites conserved in multiple organisms. We improved their algorithm in two major aspects - (i) using well-defined UTR (untranslated region) database, (ii) examining the extent of conservation inside the 3' UTR specifically. Average length in our UTR database, based on the ECgene annotation, is more than twice longer than the Ensembl. Then, TargetScan was used to identify putative binding sites. The extent of conservation varies significantly inside the 3' UTR. We used the "tight" tracks in the UCSC genome browser to select the conserved binding sites in multiple species. By combining the longer 3' UTR data, TargetScan, and tightly conserved blocks of genomic DNA, we identified 107 putative target genes with multiple binding sites conserved in multiple species, of which 85 putative targets are novel.

Keywords: microRNA, microRNA target, UTR database, comparative genomics

Introduction

MicroRNAs are small non-coding RNAs (typically 21~23 nucleotides) that play an important role in regulating post-transcriptional translation. They bind to the target mRNA by RNA base pairing, which causes target cleavage or translational repression according to the extent of complementarity. Lin-4 discovered in *C. elegans,* was the first microRNA (Lee *et al.*, 1993; Wightman *et al.*, 1993). It has multiple complementary sites in 3' UTR of lin-14. Binding of lin-4 to lin-14 leads to transition between larval stages, which suggested microRNA's role in developmental control. To date, several hundred microRNAs have been identified from various organisms and their sequences are available at the Rfam miRNA registry website (http://www.sanger.ac.uk/Software/Rfam) (Ambros *et al.*, 2003a; Griffiths-Jones, 2004).

MicroRNAs have various regulatory roles other than controlling developmental timing. For example, bantam promotes cell proliferation (Abrahante *et al.*, 2003) and Drosophila miR-14 coordinates multiple cellular responses to stress (Xu *et al.*, 2003). Recent study reports that human miR-15 and miR-16 are located in a region frequently deleted in chronic lymphocytic leukemia, which means that microRNAs may be tumor suppressors (Calin *et al.*, 2002). Functions of microRNAs are closely related to their target gene.

There are not many microRNAs with known targets in animal. Targets of lin-4, let-7, bantam, and hsa-miR-196 are experimentally verified (Abrahante *et al.*, 2003; Brennecke *et al.*, 2003; Lin *et al.*, 2003; Olsen *et al.*, 1999; Seggerson *et al.*, 2002; Yekta *et al.*, 2004). Recently we have seen many attempts to identify microRNA targets computationally in plants and animals. While simple homology search performs well for target search in plants since they usually require almost perfect complementarity (Ambros *et al.*, 2003b; Bartel *et al.*, 2003), target prediction in animal (Enright *et al.*, 2003; Kiriakidou *et al.*, 2004; Lewis *et al.*, 2003; Rajewsky *et al.*, 2004; Rhoades *et al.*, 2002; Stark *et al.*, 2003) is much more difficult because of the absence of extensive complementarity between mammalian microRNAs and their targets (Doench *et al.*, 2003; Lee *et al.*, 1993; Moss *et al.*, 1997; Olsen *et al.*, 1999; Reinhart *et al.*, 2000; Wightman *et al.*, 1993; Zeng *et al.*, 2002). It is found that the 5' core element of microRNA should be nearly perfectly complementary to the target even though partial complementarity between microRNA and its target is allowed in animal. It implies that the residue 2-8

*Corresponding author: Email sanghyuk@ewha.ac.kr,
Tel +82-2-3277-2888, Fax +82-2-3277-2384
Accepted 6 August 2005

in the microRNA is an important factor in recognizing target sites, and this microRNA seed is conserved in invertebrates and in the homologous metazoan region (Lewis *et al.*, 2003; Lim *et al.*, 2003).

Recently, Lewis *et al.* developed a robust target prediction program, TargetScan, which utilizes the seed match and multiple binding sites conserved among different species (Lewis *et al.*, 2003). They extracted 3' UTR regions from the Ensembl gene models (http://www.ensembl.org) and extended the annotated UTRs by 2kb to compensate for insufficient UTR annotation in the rat genome. TargetScan predicted 451 putative mammalian microRNA targets, and they found experimental support for 11 targets out of 15 predicted targets using the luciferase reporter assay.

However, when we examined the predicted binding sites in those 15 target genes, we found that approximately 25 % of those binding sites were not conserved in other species. Furthermore, 3' UTR region of many genes are longer than 2 kbp. We present a protocol that improves the TargetScan algorithm by constructing a better UTR database and by examining the extent of conservation inside the 3' UTR specifically.

Materials and Methods

Data preparation

MicroRNA Database: The Rfam database (http://www.sanger.ac.uk/Software/Rfam/) was used as a source of microRNA gene database (Ambros *et al.*, 2003a; Griffiths-Jones, 2004). The Rfam release 3.0 was downloaded, and 152 mature microRNAs from human were selected for target identification. Most microRNAs are identical in human, mouse, and rat.

3' UTR database: ECgene version 1.1 (hg16, mm4, rn3 for human, mouse, rat, respectively at UCSC) was used to extract the 3' UTR regions of human, mouse, and rat genomes. ECgene (Kim *et al.*, 2004; Kim *et al.*, 2005) builds gene models taking alternative splicing into consideration, and the redundant 3' UTRs from splice variants were removed from the dataset to build the non-redundant 3' UTR database. The non-redundant human 3' UTR data in the UTRdb (Pesole *et al.*, 2002) Release 17 is available at http://bighost.area.ba.cnr.it/BIG/UTRHome.

Comparative genomics data: Among several comparative genomics tracks available in the UCSC genome site, we chose the "tight" track of the conserved blocks since they were filtered with most stringent conditions. The average block size is approximately 200 bp both for human-mouse and human-rat genomes. The "net" track ignores small indels to make longer conserved blocks whose average block size is 600 bp for human-mouse comparison. The data was downloaded from the UCSC ftp site (ftp://genome.ucsc.edu/goldenPath/hg16).

MicroRNA target scan

TargetScan program available at http://genes.mit.edu/targetscan was used to find the microRNAs targets. We applied the TargetScan program for all microRNAs using our own UTR datasets from the ECgene for human, mouse, and rat genomes. We used the default parameters for TargetScan. Initially, TargetScan searches for complementary matches (so called the "seed match") for bases 2~8 of the microRNA (numbered from the 5' end). The complementary seeds are extended in both directions, and the remaining 3' portion of microRNA is optimally aligned within 35 bases flanking the target UTR seed. When the distance between adjacent seed matches is less than 20 bases, the downstream seed match is discarded as in the original TargetScan paper.

Conservation of target sites in multiple genomes

To use the comparative genomics information available at the UCSC genome center, the target positions within the UTR sequence were converted into the genomic coordinates. The gene structure should be taken into consideration since the UTR region may contain intron sequences. Determining conserved target sites is not a trivial task since there are so many putative target sites from the TargetScan and so many conserved blocks in the "tight" tracks. In an effort to reduce the number of comparisons, the binning scheme was implemented as described in the UCSC genome browser paper (Karolchik *et al.*, 2003; Kent *et al.*, 2002).

Target genes with multiple binding sites

Target genes with multiple binding sites are usually preferred. Lewis *et al.* implemented this idea as the Z score that sums contribution of multiple binding sites using the binding energies obtained from Vienna RNA package. In this study, filtering target sites using tight conservation criterion left manageable number of target sites. So we constructed a database containing all conserved target sites. A simple database query can extract target genes with multiple binding sites if necessary.

Result

Importance of cross-species conservation

Conservation of microRNA and target sequences among difference species is an important clue for their functional importance. Homology by amino acid comparison is not enough to identify microRNA target sites since we need the conservation in the 3' UTR, not in the CDS (coding sequence) region. The whole 3' UTR is not necessarily conserved in other species. We examined the conservation for 40 predicted binding sites in 14 target genes used in the luciferase reporter assay by Lewis *et al.* Using the PhyloHMM score (Siepel *et al.*, 2004a; Siepel *et al.*, 2004b) available at the UCSC genome browser, we found that only 30 binding sites had average score over 0.5. This demonstrates that the extent of conservation should be specifically examined for each binding sites in the UTR, and the comparative genomics approach is a viable method to confirm whether each binding site is conserved in multiple species.

We also performed a simulation test on the selection ratio achieved by adopting this cross-species conservation. We selected a random test set of 56 microRNAs from 152 microRNAs in the Rfam. A shuffled set was prepared by randomly permuting the microRNA sequences with their nucleotide composition unchanged. Then we ran the TargetScan to identify the binding sites using the set of longest UTR for each ECgene cluster (refer to the next section). Transcripts with multiple binding sites were 36491 and 20615 for the test set and the shuffled set, respectively. Therefore, selection ratio by multiple binding sites is 1.77:1. Examining conservation of the predicted site with PhyloHMM score over 0.9 resulted in 396 and 158 target transcripts for the two sets. This implies that the selection ratio by multiple binding sites and cross-species conservation is 2.51:1.

Construction of UTR databases

Proper 3' UTR database is necessary for microRNA target identification. Lewis *et al.* used the Ensembl database. Ensembl Human v20.34c.1 includes 22287 human genes, 16203 of those with annotated 3' UTR region. 6084 genes do not have 3' UTR region and UTR region is not long enough even for those annotated genes. To alleviate this problem, Lewis *et al.* extended the annotated 3' UTR region by 2 kbp arbitrarily in their TargetScan calculation. Even though this approach recovers UTR regions for many genes, it certainly introduces false target sites in the intergenic region.

We used the ECgene annotation that has several merits over other gene predictions. First, it includes extensive gene modeling for gene variants from alternative splicing. So UTR variation from alternative splicing or polyA can be naturally taken into consideration. Second, the 3' UTR region in the ECgene is highly reliable since it is based on EST clustering that includes ample 3' EST sequences. Furthermore, the 3' UTRs are substantially longer since they are extended by overlapping EST sequences aligned in the same orientation.

In Fig. 1, we compare the length distribution of the 3' UTR databases from the Ensembl and ECgene. We extracted 24956 transcripts with 3' UTRs in the Ensembl human using the EnsMart system (Kasprzyk *et al.*, 2004) and identified corresponding 16201 ECgene transcripts by cross-referencing with RefSeq ID. More than 40 % of the Ensembl transcripts have 3' UTRs shorter than 300 bp while only about 24% of the ECgene transcripts have less than 300 bp 3' UTRs. The average 3' UTR lengths are 549 and 1137 bp for the Ensembl and the ECgene, respectively. ECgene's UTR length is almost twice longer than corresponding Ensembl genes on average. It is also noteworthy that 2745 ECgene transcripts have 3' UTR region longer than 2 kbp and that 3' UTR region of

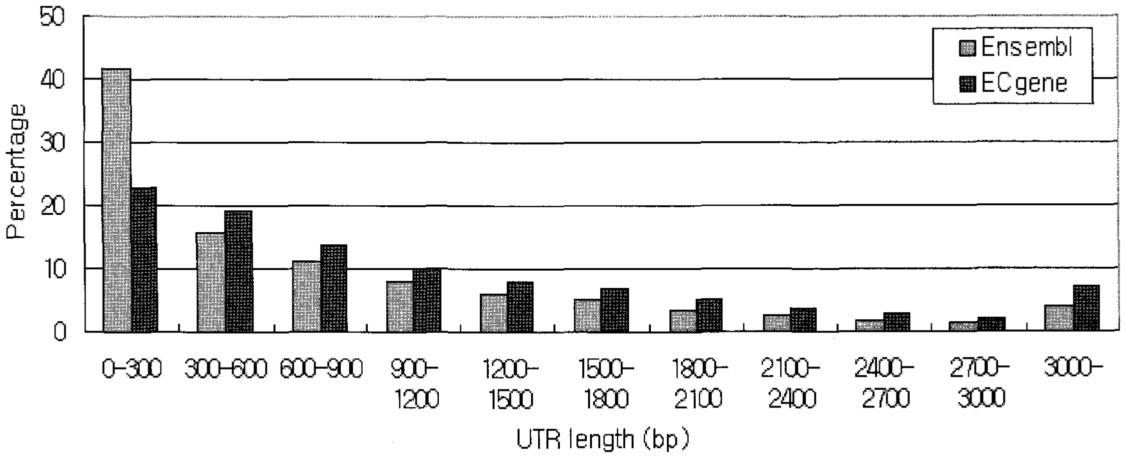

Fig. 1. UTR length distribution of the Ensembl and ECgene.

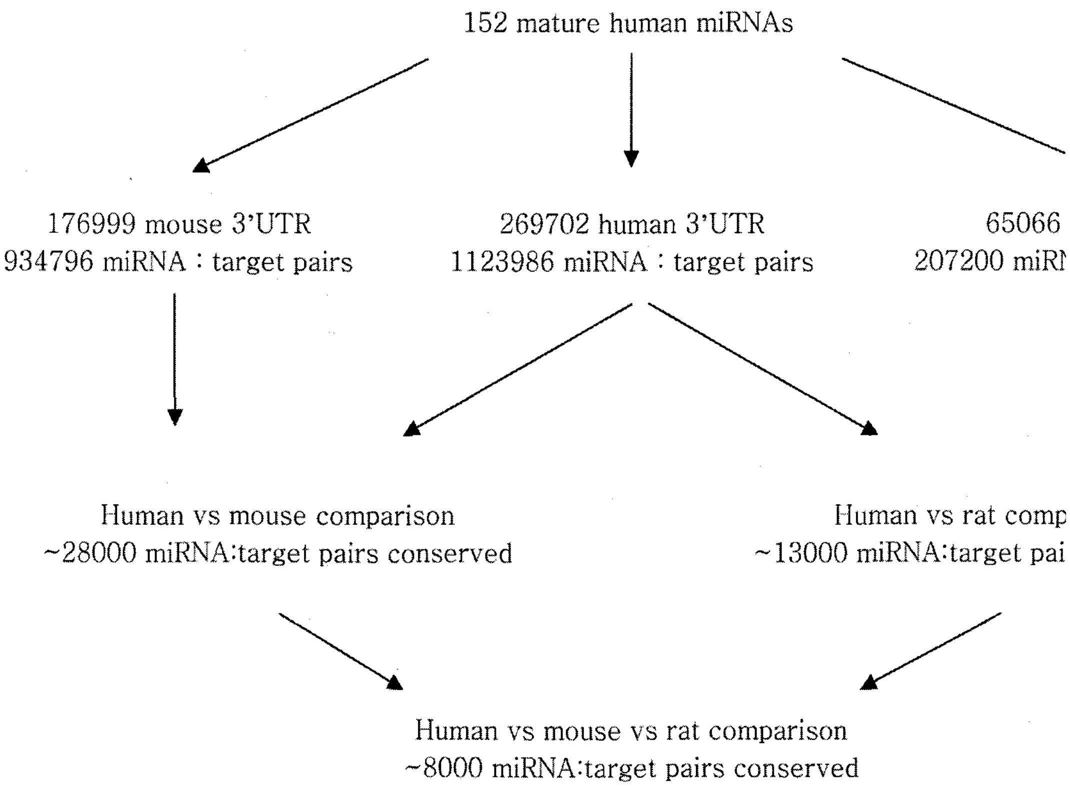
152 mature human miRNAs
176999 mouse 3'UTR
934796 miRNA : target pairs
269702 human 3'UTR
1123986 miRNA : target pairs
65066
207200 miRI
Human vs mouse comparison
~28000 miRNA:target pairs conserved
Human vs rat comp
~13000 miRNA:target pai
Human vs mouse vs rat comparison
~8000 miRNA:target pairs conserved

Fig. 2. Schematic diagram for identification of targets conserved in mammals.

1116 transcripts is longer than 3 kbp. Extending UTR region by 2 kbp in the original TargetScan work may not be effective for those genes. Similar comparison with the manually curated UTRdb (Pesole *et al.*, 2002) shows that ECgene's UTR length is longer than UTRdb by 16%. Therefore, we believe that our 3' UTR dataset should be ideal to find any charactristic features within the 3' UTR such as microRNA target sites.

Algorithmic overview

Brief overview of our strategy is summarized in Fig. 2. All human microRNAs were downloaded from the Rfam site (http://www.sanger.ac.uk/Software/Rfam/). Database of 3' UTR sequences were constructed from all ECgene (Kim *et al.*, 2004; Kim *et al.*, 2005) transcripts with annotated coding sequences. ECgene contained 269,702 distinctive 3' UTR regions for the human transcriptome. Target genes of each miRNA were predicted using the TargetScan program for our own 3' UTR datasets for human, mouse and rat. We found 1,123,986 putative target sites in the human 3' UTR data. UTR regions in mouse and rat genomes contained 934,796 and 207,200 target sites, respectively.

Comparative genomics data available at the UCSC genome center were used to find conserved target sites in three organisms. We used the "tight" track (tight subset of the best alignments obtained by BLASTZ) (Schwartz *et al.*, 2003), which satisfied the most stringent requirements. Applying the conservation criterion greatly reduced the number of target sites. We found ~28,000 target sites conserved between the human and mouse genomes, which was only 0.7% of miRNA:human target sites (1,123,986) found by TargetScan. Approximately 13,000 conserved target sites were found for the human and rat genomes. Combining two pair-wise comparisons gave ~8,000 target sites conserved in the human, mouse, and rat genomes. Selecting target genes with multiple (≥2) conserved binding sites for a specific microRNA, we found 107 target genes.

To compare our result with the original TargetScan work by Lewis *et al.*, we downloaded all target genes and the binding sites from the TargetScan website. They published 442 target genes with multiple (≥2) conserved binding sites, which were considerably larger than our 107 target genes. We shall refer to their result as "Ensembl targets" since they used the Ensembl gene annotation.

Candidate target genes

Our result is substantially different from the Ensembl targets. Only 22 out of 107 genes are found by both methods. Table 1 is the list of those common target genes from the two studies. It shows that 15 of 22 genes are targets of the same or similar microRNAs. Detailed examination, however, reveals that many genes are targets of different microRNAs even for those common target genes. The difference is clearer in Table 2 that shows the target genes found by the two methods for 10 important microRNAs. Complete list is provided as a supplementary material. We do not find many common target genes (shown in bold character), and describe factors for the difference below.

Table 1. Comparison of microRNAs targeting the same gene obtained by two methods.

Gene Symbol	ECgene ID	MicroRNAs targeting ECgene gene	microRNAs targeting Ensembl gene	Ensembl Gene ID
ZNF238	H1C27918	miR-219	miR-19b	ENSG00000179456
ZFHX1B	H2C14704	miR-141,miR-200a	miR-200b	ENSG00000169554
EPC2	H2C14989	miR-196,miR-203,miR-194	miR-196	ENSG00000135999
ACVR1	H2C15670	miR-130a,miR-130b,miR-301	miR-130,miR-130b	ENSG00000115170
RQCD1	H2C22107	miR-30a*,miR-30c,miR-30d,miR-30b, miR-30e	miR-30a,miR-30d	ENSG00000144580
CYP26B1	H2C7878	miR-93,miR-302	miR-141	ENSG00000003137
NAP1L5	H4C7703	miR-26a,miR-26b	miR-26a	ENSG00000177432
PCDHA1	H5C12441	miR-17-5p,miR-20,miR-106a,miR-185, miR-106b,miR-320	miR-181a	ENSG00000081842
HOXA5	H7C2920	miR-224	miR-19a,miR-19b	ENSG00000106004
ZHX1	H8C10857	miR-15a,miR-16,miR-103,miR-107,mi R-15b,miR-219,miR-195	miR-103,miR-107	ENSG00000189376 ENSG00000165156
FLJ20366	H8C9935	miR-19a,miR-19b,miR-130a,miR-130b ,miR-301	miR-19a,miR-19b,miR-130,miR-130b	ENSG00000147642
SURF4	H9C12513	miR-124a	miR-124a	ENSG00000188545 ENSG00000148248
PTK9	H12C4804	miR-30a*,miR-30c,miR-30d,miR-30b, miR-30e	miR-30c,miR-30d	ENSG00000151239
SLC38A2	H12C5038	miR-140	miR-26a,miR-26b	ENSG00000134294
FOXG1B	H14C1087	miR-30a*,miR-30c,miR-30d,miR-30b, miR-30e	miR-30a,miR-30b,miR-30c,miR-30d,m iR-30e,miR-200b	ENSG00000176165
CHES1	H14C7435	miR-135	miR-132,miR-135b,miR-212	ENSG00000053254
SPRED1	H15C1874	miR-17-5p,miR-20,miR-106a,miR-106b	miR-1b,miR-206	ENSG00000166068
NPTX1	H17C14064	miR-30a,miR-200a	miR-210	ENSG00000171246
FLRT3	H20C1659	miR-101	miR-101	ENSG00000125848
GGTL3	H20C3603	miR-125b,miR-125a	miR-125a,miR-125b	ENSG00000131067
SMARCA1	HXC6774	miR-9*	miR-131	ENSG00000102038

* Target genes found by both methods are listed.
* ECgene ID in version 1.1 is used.
* MicroRNAs targeting Ensembl gene are obtained from the TargetScan website (http://genes.mit.edu/targetscan/)

Table 2. Target genes for 10 selected microRNAs.

microRNA	ECgene	Ensembl
let-7a	BZW1	PUNC, HIC2, NR6A1(GCNF), ADAMTS15, C10orf6, LIMK2, BZW1, APEX1, SFMBT1
let-7b	BZW1	PUNC, NR6A1(GCNF), HIC2, ADAMTS15, LIMK2, C10orf6, APEX1, BZW1, NME4, FBXO32, SFMBT1
miR-15a	ZHX1, BCL2L2	ARHGDIA, IHPK1, DEDD, GCAT, MFN2, C1orf37
miR-16	ZHX1, BCL2L2	IHPK1, DEDD, GCAT, MFN2, C1orf37
miR-23a	DNAJC6, PDE4B, MAP4K4, MO25, BTBD14A	POU4F2, CXCL12, ZNF292, UBE2D3, PPIF, NEK6, FBXO32
miR-26a	EIF4G2, DRLM	SLC38A2, MADH1, PRKWNK3, NAB1, PELI2
miR-30a	NPTX1	TNRC6, FOXG1C, FOXG1B, GLI2, GNAI2, BCL9, RQCD1, SEC24A, STIM2, LHX1, RARG, LIN28
miR-124a	SURF4	ARHQ, SURF4, PMX1, LPIN1, OSBPL3, TCF3, MITF, PEA15, ELOVL5, CD164, SDFR1, SEMA6C, STAT3, HMG2L1, ANGPT1
miR-181a	HOXC8, TRIP15, BTBD3, CGI-72	ESM1, PERQ1, HOXA11, PCDHA family, PCDHAC1, PCDHAC2, BCL6B
miR-196	HOXC8, EPC2	HOXA7, NR6A1, CALM1, CALM2, CALM3

* Common target genes are shown in bold character.
* Genes with a valid HUGO gene symbol are included.

The most influential factor is the way to determine whether the target site is conserved in other species. Lewis *et al.* used the homologous gene information in the Ensembl annotation system (Birney *et al.*, 2004). Ensembl homologues are obtained by comparing the amino acid sequence of proteins (Clamp *et al.*, 2003). Even though this is perfectly valid for comparing coding regions, it does not guarantee that the UTR regions are conserved. The "Conservation" track (Siepel *et al.*, 2004a) in Fig. 3 shows the conservation level obtained by comparing five organisms (human, chimp, mouse, rat, and chicken). It can be immediately recognized that the extent of sequence conservation varies dramatically inside the 3' UTR region. Our method uses the "tight" track that includes conserved blocks of high quality only as can be seen in Fig. 3. This is the major reason why the number of target genes is substantially smaller in the ECgene targets. All of our target sites are highly likely to be conserved in other species, which reduces risk of being false positives.

Another big difference is the gene model and the resultant UTR. ECgene's UTR is significantly longer than other gene models since its algorithm tries to extend the UTR regions if it finds any overlapping ESTs with the same direction. Detailed comparison is given in the next section. Furthermore, alternative splicing is extensively modeled in the ECgene (Kim *et al.*, 2004; Kim *et al.*, 2005), which greatly expands the number of transcripts. Current Ensembl database (Ensembl Human v20.34c.1) includes 35685 transcripts from 23758 genes. Approximately 70% (16534) of Ensembl have single transcript and remaining 7224 genes give 19151 transcripts, implying 2.65 isoforms per gene if the gene is alternatively spliced. However, current knowledge implies that ~70% of human genes are alternatively spliced and that ~3.5 isoforms exist for alternatively spliced genes. Therefore transcripts in the Ensembl do not reflect diversity from alternative splicing adequately. Our UTR database based on ECgene has 269702 distinct UTR regions. Even though it may include many false positives, its coverage should be most complete. The difference is explored in the following cases in detail.

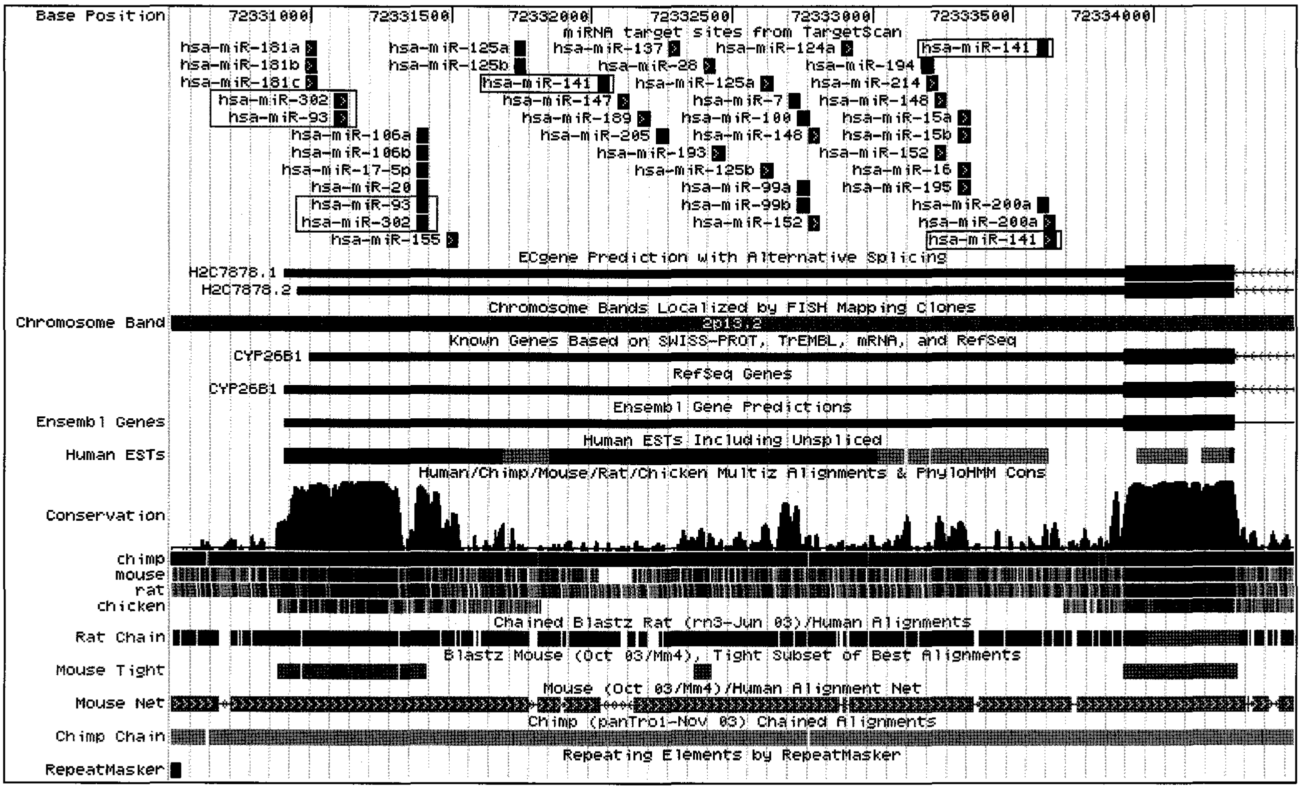

Fig. 3. Target binding sites and sequence conservation in CYP26B1. The first track in black color shows all target binding sites found by TargetScan. MicroRNAs with multiple binding sites are shown in colored boxes. The gene is on the antisense strand (3' end on the left side) as indicated by arrows in the intron. The genomic region is mostly the last exon, where the CDS is tall and the 3' UTR region is short. Lower half shows several tracks for comparative genomics. Conservation track in black clearly shows that the binding sites for hsa-miR-141 are not conserved in multiple species. The track labeled as "Mouse Tight" is the most stringent collection of the conserved regions between human and mouse.

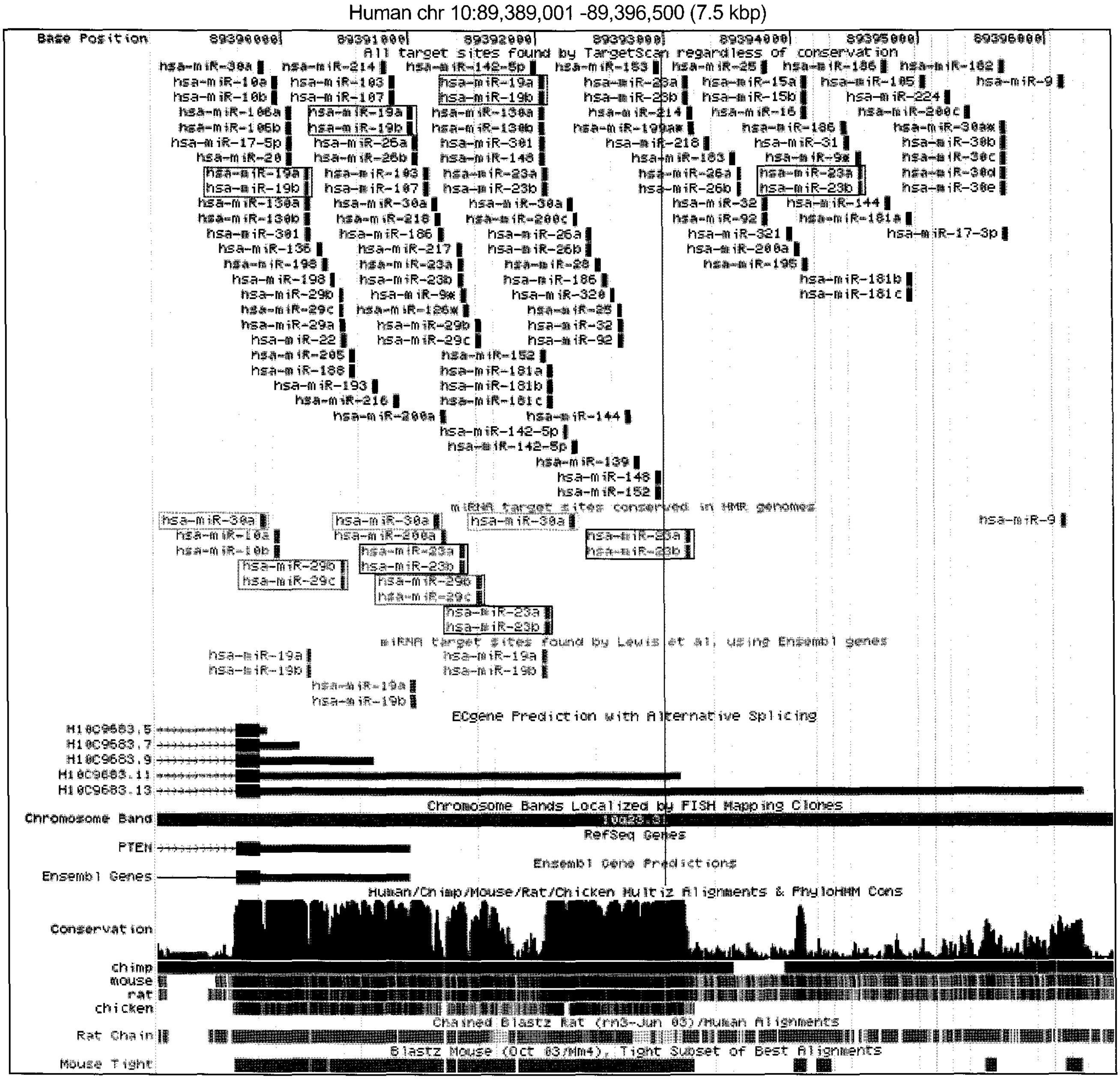
Human chr 10:89,389,001 -89,396,500 (7.5 kbp)
Base Position
89390000
89391000
89392000
89393000
89394000
89395000
89396000
All target sites found by TargetScan regardless of conservation
miRNA target sites conserved in HMR genomes
miRNA target sites found by Lewis et al, using Ensembl genes
ECgene Prediction with Alternative Splicing
H10C9683.5
H10C9683.7
H10C9683.9
H10C9683.11
H10C9683.13
Chromosome Bands Localized by FISH Mapping Clones
Chromosome Band
10q23.31
RefSeq Genes
PTEN
Ensembl Gene Predictions
Ensembl Genes
Human/Chimp/Mouse/Rat/Chicken Multiz Alignments & PhyloHMM Cons
Conservation
chimp
mouse
rat
chicken
Chained Blastz Rat (rn3-Jun 03)/Human Alignments
Rat Chain
Blastz Mouse (Oct 03/mm4), Tight Subset of Best Alignments
Mouse Tight

Fig. 4. Genome browser for the 3' UTR of PTEN_human. Binding sites conserved in multiple species are shown in colored boxes. Long vertical line indicates the end of 2 kbp extended region for 3' UTR of the Ensembl transcript. The violet track shows the binding sites conserved in human, mouse, and rat genomes. The red track is the target sites found by Lewis *et al.* Other information is comparable to

Case study 1: CYP26B1

We observed in Tables 1 and 2 that target genes are much different between two methods. For example, CYP26B1 (Nelson, 1999), a member of the cytochrome P450 superfamily of enzymes, is target of miR-93 and miR-302 in our study, whereas miR-141 is the binding microRNA in the original TargetScan study.

Fig. 3 summarizes relevant information for CYP26B1 in the genome browser. It shows the gene structures, binding sites, and comparative genomics data. All target sites found by the TargetScan program are listed without filtering by conservation or multiple binding sites. The 3' UTR regions are more or less the same between the ECgene, Ensembl gene, and RefSeq gene tracks. Both miR-93 and miR-302 have two identical binding sites (shown in the red boxes). Several tracks are available for comparative genomics. The "conservation" track (Human/Chimp/Mouse/Rat/Chicken Multiz Alignments & PhyloHMM Cons) is especially useful (Siepel *et al.*, 2004b). It can be immediately seen that we have two large conserved blocks - one for the last exon and the other near the end of 3' UTR region. The "mouse tight" track is collection of most highly conserved blocks. Other tracks such as "chain" and "net" are conservation with less stringent conditions. We observe that two binding sites of miR-93 and miR-302 appear within the highly conserved genomic locus.

The Ensembl target, hsa-miR-141, has three binding sites (shown in boxes) for this gene, which is advantageous over miR-93 and miR-302 with two binding sites. However, the conservation level for these target sites is much lower in the "conservation" track, and they do not qualify for conserved blocks in the mouse tight track. Even though targets with multiple binding sites are generally preferred, the mechanism underlying the cooperativeness between the binding sites is not established well. Conservation in other species should be a clear evidence for its functional importance. Therefore, CYP26B1 would not be a good target for hsa-miR-141 since its binding sites are not conserved in high standard. This example shows that finding target genes for microRNAs needs subtle adjustment of various aspects such as multiple binding sites and conservation among different species.

Case study 2: PTEN

PTEN (phosphatase and tensin homolog) is an important gene involved in regulation of the AKT1 signaling pathway. Mutations of PTEN are found in a large number of cancers and it is a potential tumor suppressor gene (Di Cristofano *et al.*, 2000). Lewis *et al.* identified PTEN as a target of two microRNAs - hsa-miR-19a and hsa-miR-19b. Our method did not find any microRNA targeting this gene. The main reason was that the PTEN annotation in rat did not have UTR at all. Lewis *et al.* cleverly extended the 3' UTR region by 2 kbp to alleviate this problem.

Fig. 4 shows the 3' UTR of human PTEN with predicted target sites. The 3' UTR is unusually long – 6.5 kbp and 1.2 kbp in the ECgene and Ensembl, respectively. Even though the human PTEN has three binding sites as Lewis *et al.* found, only the first two binding sites are found in the mouse PTEN and the rat PTEN has only one binding site. So hsa-miR-19a and hsa-miR-19b do not have multiple conserved binding sites in PTEN.

Instead, we identified many other conserved binding sites as shown in Fig. 4. The major hurdle in identifying conserved targets is the poor annotation in the rat genome. The UTR in the mouse PTEN is 3.3 kbp long, but there are many neighboring EST transcripts and it is expected to be as long as the UTR of human PTEN. So, we extended the UTRs of PTEN in mouse and rat to be 6.5 kbp comparable to the human PTEN. The number of binding sites found by TargetScan is shown in Table 3. We found additional binding sites for five microRNAs (hsa-miR-30a, hsa-miR-23a, hsa-miR-23b, hsa-miR-29b, hsa-miR-29c). Furthermore, four additional microRNAs (hsa-miR-200a, hsa-miR-181a, hsa-miR-181b, hsa-miR-181c) were identified to have multiple conserved binding sites for PTEN.

Hsa-miR-23a and hsa-miR-23b have four binding sites in human as shown in Fig. 4 and Table 3. The fourth binding site does not belong to the conserved block as can be seen in the conservation track. Three other binding sites appear inside the conserved region. However, the third binding site is located outside the 2 kbp extended region of Ensembl UTR (indicated as long vertical lines in brown color in Fig. 4). So, extending by 2 kbp would not be enough for genes with long 3' UTR.

Discussion

Comparison of our method with the original TargetScan program suggests many important aspects in target prediction of microRNAs. First, using properly annotated 3' UTR database is essential. Extending 3' UTR by an arbitrary length is not desirable since it increases false positives by introducing target sites outside the real UTR. Extension of 2 kbp was not enough for the case study of PTEN. We need a compromise between sensitivity and specificity. Selecting 3' UTR of reliable transcripts in the

Table 3. Number of target sites in the 3' UTR of PTEN for various microRNAs

MicroRNA	Number of target binding sites in 3' UTR of PTEN		
	Human	Mouse	Rat
hsa-miR-19a, b	3	2	1
hsa-miR-30a	**3**	**2**	**2**
hsa-miR-23a, b	**4**	**2**	**4**
hsa-miR-29b, c	**2**	**2**	**3**
hsa-miR-200a	**2**	**3**	**3**
hsa-miR-181a, b, c	**2**	**3**	**3**
hsa-miR-186	4	1	1
hsa-miR-26a, b	3	1	1
hsa-miR-142-5p	3	1	1
hsa-miR-9	1	1	1
hsa-miR-10a, b	1	1	1

* MicroRNAs in **bold** character have multiple binding sites in human, mouse, and rat.
* The 3' UTR is extended to be 6.5 kbp for all cases.

ECgene or Acembly genes (D. Therry-Mieg *et al.* http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly) database would be the best choice for human and mouse genomes at this point. However, this is not true for rat which has only ~0.6 million ESTs in the dbEST, compared with 5.64, 4.18 million ESTs for human and mouse, respectively. Proper annotation of 3' UTR is impossible in rat due to the draft quality of the rat genome and insufficient EST data. Their effect on target prediction of microRNAs is obvious as can be seen in Fig. 2. Number of rat transcripts is approximately 1/3 of human or mouse, and the number of microRNA:target pairs are reduced to 1/5 in rat. It might be better to compare just human and mouse genomes, waiting for the rat genome map to be more complete. This would increase the number of target genes substantially. Alternatively, one may extend the 3' UTR of rat by the amount inferred by human and mouse homologs, not beyond the gap position.

Another important aspect is the conservation of target sites in different organisms. As seen in Fig. 3 and 4, the extent of conservation varies significantly inside the 3' UTR, which is not taken into consideration in previous studies. It would be better to count the target binding sites within the tightly conserved blocks only. Two comparative genomics tracks-PhyloHMM and Blastz tight - in the UCSC genome browser should be good candidates (Schwartz *et al.*, 2003; Siepel *et al.*, 2004a). We used the tight tracks for convenience since it provides pairwise comparison of genomes.

Recent study of Yekta *et al.* identified HOXB8 as the target of hsa-miR-196 (Yekta *et al.*, 2004). The target site is nearly perfectly complementary to the microRNA. TargetScan allows target binding with mismatches and bulges, producing target sites of 42-43 nucleotides long. Interestingly, TargetScan could not locate the target site for hsa-miR-196 because the target site has one mismatch in the 5' seed region (2-8 bp of microRNA). This suggests that minor mismatches in the seed region should be allowed in search of microRNA targets.

Acknowledgements

We are grateful to the UCSC genome center for the comparative genomics data and to C. B. Burge and coworkers for TargetScan program. This work was supported by the Ministry of Science and Technology of Korea through the bioinformatics research program of MOST NRDP (M1-0217-00-0027) and the Korea Science and Engineering Foundation through the center for cell signaling research at Ewha Womans University.

References

Abrahante, J.E., Daul, A.L., Li, M., Volk, M.L., Tennessen, J.M., Miller, E.A., and Rougvie, A.E. (2003). The Caenorhabditis elegans hunchback-like gene lin-57/hbl-1 controls developmental time and is regulated by microRNAs. *Dev. Cell* 4, 625-637.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M., Matzke, M., Ruvkun, G., and Tuschl, T. (2003a). A uniform system for microRNA annotation. *RNA* 9, 277-279.

Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T., and Jewell, D. (2003b). MicroRNAs and other tiny endogenous RNAs in C. elegans. *Curr. Biol.* 13, 807-818.

Bartel, B. and Bartel, D.P. (2003). MicroRNAs: at the root of plant development? *Plant Physiol.* 132, 709-717.

Birney, E., Andrews, D., Bevan, P., Caccamo, M., Chen, Y., Clarke, L., *et al.* (2004). Ensembl 2004. *Nucleic Acids Res.* 32, D468-D470.

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.

Calin, G.A., Dumitru, C.D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., Aldler, H., Rattan, S., Keating, M., Rai, K., Rassenti, L., Kipps, T., Negrini, M., Bullrich, F., and Croce, C.M. (2002). Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc. Natl. Acad. Sci. USA* 99, 15524-15529.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., *et al.* (2003). Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.* 31, 38-42.

Di Cristofano, A. and Pandolfi, P.P. (2000). The multiple roles of PTEN in tumor suppression. *Cell* 100, 387-390.

Doench, J.G., Petersen, C.P., and Sharp, P.A. (2003). siRNAs can function as miRNAs. *Genes Dev.* 17, 438-442.

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

Griffiths-Jones, S. (2004). The microRNA Registry. *Nucleic Acids Res.* 32, D109-D111.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D., and Kent, W.J. University of California Santa Cruz. (2003). The UCSC Genome Browser Database. *Nucleic Acids Res.* 31, 51-54.

Kasprzyk, A., Keefe, D., Smedley, D., London, D., Spooner, W., Melsopp, C., Hammond, M., Rocca-Serra, P., Cox, T., and Birney, E. (2004). EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* 14, 160-169.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* 12, 996-1006.

Kim, N., Shin, S., and Lee, S. (2004). ASmodeler: Gene modeling of alternative splicing from genomic alignment of mRNA, EST, and protein sequences. *Nucleic Acids Res.* 32, W181-W186.

Kim, P., Kim, N., Lee, Y., Kim, B., Shin, Y., and Lee, S. (2005). ECgene: genome annotation for alternative splicing. *Nucleic Acids Res.* 33, D75-D79.

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.

Lee, R.C., Feinbaum, R.L., and Ambros, V. (1993). The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 75, 843-854.

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.

Lim, L.P., Lau, N.C., Weinstein, E.G.., Abdelhakim, A., Yekta, S., Rhoades, M.W., Burge, C.B., and Bartel, D.P. (2003). The microRNAs of Caenorhabditis elegans. *Genes Dev.* 17, 991-1008.

Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E., and Slack, F.J. (2003). The C elegans hunchback homolog, hbl-1, controls temporal patterning and is a probable microRNA target. *Dev. Cell* 4, 639-650.

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.

Nelson, D.R. (1999). A second CYP26 P450 in humans and zebrafish: CYP26B1. *Arch. Biochem. Biophys.* 371, 345-347.

Olsen, P.H. and Ambros, V. (1999). The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* 216, 671-680.

Pesole, G., Liuni, S., Grillo, G., *et al.* (2002). UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. Update 2002. *Nucleic Acids Res.* 30, 335-340.

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

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.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. (2002). Prediction of plant microRNA targets. *Cell* 110, 513-520.

Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Res.* 13, 103-107.

Seggerson, K., Tang, L., and Moss, E.G.. (2002). Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation. *Dev. Biol.* 243, 215-225.

Siepel, A. and Haussler, D. (2004a). Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* 11, 413-428.

Siepel, A. and Haussler, D. (2004b). Phylogenetic estimation of context-dependent substitution rates by maximum likelihood. *Mol. Biol. Evol.* 21, 468-488.

Stark, A., Brennecke, J., Russell, R.B., and Cohen, S.M. (2003). Identification of Drosophila MicroRNA Targets. *PLoS. Biol.* 1, E60.

Wightman, B., Ha, I., and Ruvkun, G.. (1993). Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* 75, 855-862.

Xu, P., Vernooy, S.Y., Guo, M., and Hay, B.A. (2003). The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr. Biol.* 13, 790-795.

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

Zeng, Y., Wagner, E.J., and Cullen, B.R. (2002). Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. *Mol. Cell* 9, 1327-1333.