COMPUTATIONAL DISCOVERY OF ANIMAL SMALL RNA GENES AND TARGETS

Inauguraldissertation

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Dimosthenis Gaidatzis

aus Griechenland

Basel, 2007

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät der Universität Basel auf Antrag der Professoren Mihaela Zavolan und Félix Naef.

Prof.Mihaela Zavolan(Referent)

Prof.Félix Naef(Koreferent)

Basel, den 11.Dezember 2007

Prof. Dr. Hans-Peter Hauri

Dekan

ABSTRACT

COMPUTATIONAL DISCOVERY OF ANIMAL SMALL RNA GENES AND TARGETS

Dimosthenis Gaidatzis

Biozentrum, University of Basel

Swiss Institute of Bioinformatics

Though recently discovered, small RNAs appear to play a wealth of regulatory roles, being involved in degradation of target mRNAs, translation silencing of target genes, chromatin remodeling and transposon silencing. Presented here are the computational tools that I developed to annotate and characterize small RNA genes and to identify their targets. One of these tools is oligomap, a novel software for fast and exhaustive identification of nearly-perfect matches of small RNAs in sequence databases. Oligomap is part of an automated annotation pipeline used in our laboratory to annotate small RNA sequences. The application of these tools to samples of small RNAs obtained from mouse and human germ cells together with subsequent computational analyses lead to the discovery of a new class of small RNAs which are now called *piRNAs*. The computational analysis revealed that piRNAs have a strong uridine preference at their 5' end, that unlike miRNAs, piRNAs are not excised from fold-back precursors but rather from long primary transcripts, and that the genome organization of their genes is conserved between human and mouse even though piRNAs on the sequence level are poorly conserved. In vertebrates, the most studied class of small regulatory RNAs are the miRNAs which bind to mRNAs and block translation. A computational framework is introduced to identify miRNA targets in mammals, flies, worms and fish. The method uses extensive cross species conservation information to predict miRNA binding sites that are under evolutionary pressure. A downstream analysis of predicted miRNA targets revealed novel properties of miRNA target sites, one of which is a positional bias of miRNA target sites in long mammalian 3' untranslated regions. Intersection of our predictions with biochemical pathway annotation data suggested novel functions for some of the miRNAs. To gain further insights into the mechanism of miRNA targeting, I studied microarray data obtained in siRNA experiments. SiRNAs have been shown to produce off-targets that resemble miRNA targets. This analysis suggests the presence of additional determinants of miRNA target site functionality (beyond complementarity between the miRNA 5' end and the target) in the close vicinity (about 150 nucleotides) of the miRNA-complementary site. Finally, as part of a study aiming to reduce siRNA off-target effects by introducing chemical modifications in the siRNA, I performed microarray data analysis of siRNA transfection experiments. Presented are the methods used to quantify off-target activity of siRNAs carrying different types of chemical modifications. The analysis revealed that off-targets caused by the passenger strand of the siRNA can be reduced by 5'-O-methylation.

ACKNOWLEDGMENTS

I gratefully acknowledge Mihaela Zavolan, Erik van Nimwegen, Thomas Tuschl, Gunter Meister, Jean Hausser, Alain Sewer, Philipp Berninger, Alexei Aravin and Po Yu (Jeff) Chen for support and fruitful collaborations.

Contents

Chapter 1

Annotation and characterization of small RNA sequence libraries

1.1 Introduction

Cloning and sequencing is the method of choice for small regulatory RNA identification. Using deep sequencing technologies one can now obtain up to a billion nucleotides – and tens of millions of small RNAs – from a single library. Careful computational analyses of such libraries enabled the discovery of miRNAs, rasiRNAs, piRNAs, and 21U RNAs. Given the large number of sequences that can be obtained from each individual sample, deep sequencing may soon become an alternative to oligonucleotide microarray technology for mRNA expression profiling.

Though recently discovered, small RNAs appear to play a wealth of regulatory roles, ranging from degradation of target mRNA [1, 2], translation silencing of target mRNA [3–5], chromatin remodeling [6, 7] and transposon silencing [8–10]. In vertebrates, the most studied class of small regulatory RNAs are the microRNAs (miRNAs), which are produced from hairpin precursors by the Dicer endonuclease [3–5] to block the translation of target mRNAs [11]. The discovery of the let-7 miRNA, which is perfectly conserved in sequence from worm to man [12], sparked a great interest in the identification of additional miRNAs as well as of other regulatory RNAs. The group of Tom Tuschl (Rockefeller University) developed a protocol for isolating miRNAs which typically yields $80-90\%$ miRNAs in a given sample of small RNAs [13,14], and used it to collect small RNA expression profiles from hundreds of mammalian samples. Based on this data, an atlas of miRNA expression profiles in a large number of mammalian tissues [15] was constructed. In parallel, high-throughput pyrosequencing [16] or sequencing-by-synthesis [17] technologies are being developed to deliver up to a billion nucleotides in a run. With millions of miRNA sequences from a single sample, one can obtain a very fine resolution picture of miRNA expression.

As is generally the case with high-throughput data, fast and accurate computational analysis methods are needed to uncover the information contained in these large datasets. One of the main goals of my work was to develop computational methods for the identification and characterization of novel regulatory RNAs. Here I will present one of my projects that addressed this topic. It concerns a set of analyses that I performed in order to identify and characterize a novel class of small regulatory RNAs that specifically associate with proteins of the Piwi family that are expressed in germ cells [18].

1.2 Automated annotation of small RNAs

The protocol for small RNA sequencing is sketched in Figure 1.1. Total RNA is size-separated to extract sequences of the appropriate size (roughly 22 nucleotides for miRNAs, 25-35 for piRNAs, etc.), which are subjected to adaptor ligation using a procedure that takes advantage of the presence of a 5' phosphate and a 3' hydroxyl group in the RNase III products [14]. The resulting sequences are concatenated, ligated into the T vector, cloned and sequenced. The first computational step is to retrieve the sequence of the small RNAs from the sequenced concatamers. This is accomplished by mapping the adaptors to the concatamer sequences. The subsequences of a concatamer that are found between matches to 5' and 3' adaptors in the correct configuration are extracted as small RNAs.

The first aim of the analysis of a large-scale small RNA dataset is to identify all sequences whose function is already known. Since many genomes have been now sequenced and annotated to a large extent, one can frequently infer the function of a small RNA from the annotation of the genomic region to which the small RNA maps. This approach of course fails when the genome assembly or the genome annotation

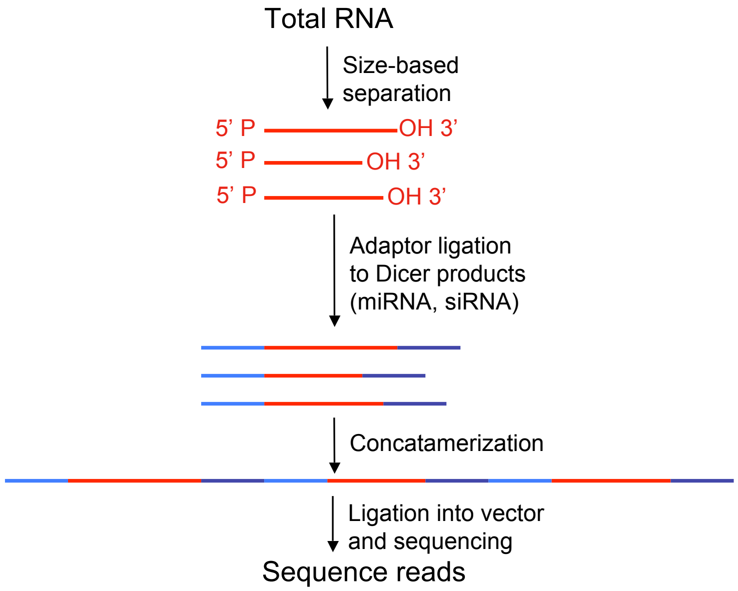


Figure 1.1: Protocol for small RNA sequencing.

are incomplete or incorrect. For instance, the annotation of small RNAs derived from ribosomal RNA cannot be readily done based on the genome annotation because the rRNA repeat unit, though available in the Genbank database (U13369 for human and BK000964 for mouse), is not present in its entirety in the current assemblies of the human and mouse genomes. Another example is the cluster of mouse embryonic miRNAs (mmu-mir-290 to mmu-mir-295) which is absent from the current assembly of the mouse genome, but was present in a previous assembly [19]. For this reason we used in our study both the genome annotation as well as mappings of the small RNAs to transcripts with known function to functionally annotate small RNAs. We downloaded the genome sequence of the species from which the small RNAs have been cloned from the UCSC repository (http://genome.cse.ucsc.edu), from which we also obtain the annotation of repeat elements in the genome. As sources of transcripts of known function we used the following resources:

- miRNA - ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/CURRENT/hairpin.fa.gz
- rRNA - Genbank sequence search using human/mouse/rat as species and rRNA

as "Molecule type" to filter the records

- tRNA - http://lowelab.ucsc.edu/GtRNAdb/
- sn- and sno-RNA - Genbank sequence search filtering for the appropriate "Feature Key"
- mRNA - Genbank sequence search using human/mouse/rat as species and mRNA as "Molecule type" to filter the records.

After compiling these data, we can proceed to identify those small RNAs whose function is already known. We achieved through the following steps:

1. Small RNAs are mapped to genome using oligomap ($0-/1-$error matches,see Chapter 2) and WU-BLAST (matches with ≥ 2 errors).
2. For each small RNA the locus/loci with minimum number of errors (mismatch, insertion, deletions) in the small RNA-to-genome mapping is/are identified.
3. Too distant mappings ($< 92.5\%$ identity) are filtered out.
4. Small RNAs are mapped to annotated sequences using oligomap ($0-/1-$error matches), WU-BLAST (matches with ≥ 2 errors).
5. For each small RNA the sequences with minimum number of errors (mismatch, insertion, deletions) in the small RNA-to-annotated sequence mapping are identified.
6. Too distant mappings $< 92.5\%$ identity are filtered out.
7. A functional category is assigned to each small RNA based on all its best mappings.

Many small RNAs map unambiguously to sequences from one single functional category, and their origin can therefore be easily determined. There typically are also small RNAs that map equally well to sequences with different function, such as for instance tRNA and genomic repeat. For these, we choose what we consider the most likely annotation based roughly on the abundance of various types of sequences in the cell, namely rRNA > tRNA > sn/sno-RNA > miRNA > piRNA > repeat > mRNA.

Using this annotation procedure, we have determined that RNAs that immunoprecipitate with the Mili protein of the Piwi family are depleted in miRNAs or other known RNAs, but are enriched in sequences of unknown function, whose average length is 26-30 nucleotides, i.e. are longer than miRNAs, and that, similarly to miRNAs had a very strong U-bias at the first position. We therefore set to characterize these sequences, as described in the following section.

1.3 A novel class of small RNAs bind to MILI protein in mouse testes

Parts of this section have been published in [18].

Alexei Aravin[1,9,10], Dimos Gaidatzis[2,9], Sbastien Pfeffer[1], Mariana Lagos-Quintana[1], Pablo Landgraf[1], Nicola Iovino, Patricia Morris[3], Michael J. Brownstein[4], Satomi Kuramochi-Miyagawa[5], Toru Nakano[5], Minchen Chien[6], James J. Russo[6], Jingyue Ju[6],[7], Robert Sheridan[8], Chris Sander[8], Mihaela Zavolan[2,*],[*] & Thomas Tuschl[1,*]

[1]Howard Hughes Medical Institute, Laboratory of RNA Molecular Biology, The Rockefeller University, 1230 York Avenue, Box 186, New York, NY 10021, USA.

[2]Biozentrum, Universität Basel, Klingelbergstr 50-70, CH-4056 Basel, Switzerland. Swiss Institute of Bioinformatics

[3]Population Council, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA.

[4]J. Craig Venter Institute, Functional Genomics, 9704 Medical Center Drive, Rockville, MD 20850, USA.

[5]Department of Pathology, Medical School, Graduate School of Frontier Biosciences, Osaka University, Yamada-oka 2-2 Suita, Osaka 565-0871, Japan.

[6]Columbia Genome Center, Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA.

[7]Department of Chemical Engineering, Columbia University, 500 West 120 Street, New York, NY 10027, USA. [8]Computational Biology Center, Memorial Sloan-Kettering

Cancer Center, New York, NY 10021, USA.
[9]These authors contributed equally to this work.
[10]Present address: Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA.
*Corresponding authors

1.3.1 Abstract

Small RNAs bound to Argonaute proteins recognize partially or fully complementary nucleic acid targets in diverse gene–silencing processes [20–23]. A subgroup of the Argonaute proteins–known as the Piwi family [24]-is required for germ– and stem–cell development in invertebrates [25,26], and two Piwi members–MILI and MIWI–are essential for spermatogenesis in mouse [27, 28]. Here we describe a new class of small RNAs that bind to MILI in mouse male germ cells, where they accumulate at the onset of meiosis. The sequences of the over 1,000 identified unique molecules share a strong preference for a 5' uridine, but otherwise cannot be readily classified into sequence families. Genomic mapping of these small RNAs reveals a limited number of clusters, suggesting that these RNAs are processed from long primary transcripts. The small RNAs are 26–31 nucleotides (nt) in length–clearly distinct from the 21–23 nt of microRNAs (miRNAs) or short interfering RNAs (siRNAs)–and we refer to them as Piwi–interacting RNAs or piRNAs. Orthologous human chromosomal regions also give rise to small RNAs with the characteristics of piRNAs, but the cloned sequences are distinct. The identification of this new class of small RNAs provides an important starting point to determine the molecular function of Piwi proteins in mammalian spermatogenesis.

1.3.2 MILI–immunoprecipitation from testis lysate of adult mice

MILI–containing ribonucleoprotein complexes were immunoprecipitated from testis lysate of adult mice and purified the associated RNAs. Although 5' ^{32}P–labelling of the isolated molecules revealed a distinct population of RNAs 26–28 nt in length,

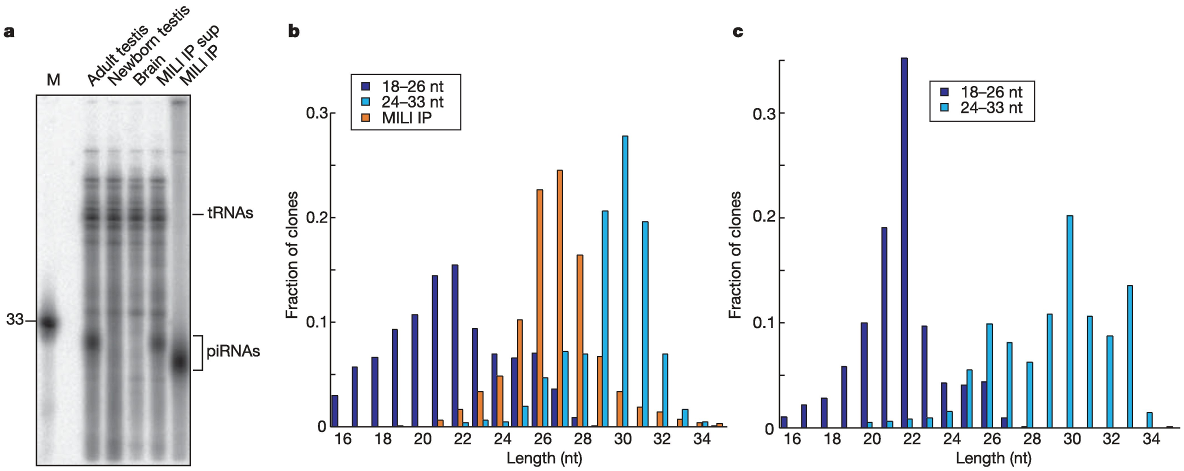


Figure 1.2: MILI associates with 26–28 nt RNAs. a, 5’ ^{32}P–labelling of total RNA isolated from testes of adult or 10–day–old mice and adult brain. Adult testis reveals an abundant 29– to 31–nt small RNA fraction. MILI–immunoprecipitated (IP) RNAs are predominantly 26– to 28–nt while 29– to 31–nt RNAs remain in the supernatant (sup). Size and mobility of oligoribonucleotide marker is indicated on the left. b, Size distribution of small RNAs cloned from adult mouse testis total RNA of 18– to 26–nt (dark blue bars) and 24– to 33–nt (light blue bars), and MILI–immunoprecipitated RNA (orange bars). c, Size distribution of small RNAs cloned from human testis total RNA of 18– to 26–nt (dark blue bars) and 24– to 33–nt (light blue bars)

total testis RNA from fertile adults–but not 10–day–old mice-showed a prominent RNA species of approximately 30 nt in length (Figure 1.2a). Labelling of the RNA remaining in the supernatant of the MILI immunoprecipitation showed that the 30–nt fraction was intact, and that the 26–28–nt MILI–interacting RNAs were unlikely to represent degradation products of the 30–nt–long RNAs. To determine the identity of the different small–RNA size populations, the MILI–interacting small RNAs, as well as small RNAs from testis ranging in size between 18–26 nt and 24–33 nt were cloned and sequenced, and analyzed as described below.

The over 15,000 sequences identified in three small–RNA libraries suggested three distinct size populations of small RNAs. The 18–26–nt fractionated library showed a peak at 21 nt and 22 nt, corresponding to miRNAs, and also revealed a 26–nt shoulder (Figure 1.2b). The 24–33–nt fraction showed a bimodal distribution with a strong peak at 29–31 nt, corresponding to the small RNAs detected by 5'-labelling of total testis RNA, and a small peak at 26–28 nt. MILI–interacting small RNAs demonstrated a unimodal length–distribution, with a peak at 26–28 nt. The 27–nt shoulders in the size distribution profiles of the smaller and larger size fractions thus probably represent the MILI–interacting subpopulation. Whereas ∼60% of small RNA clones from the 18–26–nt library can be annotated as miRNAs and degradation products of abundant cellular RNAs, more than 80% of the sequences in the MILI immunoprecipitate and the 24–33–nt libraries did not derive from known transcripts or genomic repeats (Table 1.1).

1.3.3 Characterization of piRNAs

Sequence analysis indicated that 85% of the MILI–interacting RNAs and 88% of the 24–33–nt RNAs contain a 5' uridine residue (Table 1.1, Figure 1.3). The bias for 5' uridine is one of the characteristics of miRNAs and repeat–associated siRNAs (rasiRNAs) produced from double–stranded RNA (dsRNA) precursors by RNase III enzymes [29–31]. The majority of mouse small RNA sequences (92.9% and 97.7%, respectively) obtained from libraries of 24–33–nt testis and MILI immunoprecipitation were cloned only once. Genomic mapping of cloned sequences showed that less than 15% of clones in both libraries are derived from annotated repetitive regions (Tables 1.1,1.2). Further more, less than 3% of the sequences match the genome more than ten times. The relative frequency of sequences matching different types of repeats roughly corresponds to the relative proportion of those repeat elements in the genome.

Clustering the genomic loci corresponding to the more than 1,500 MILI–interacting sequences, on the basis of a maximum distance of 15 kilobases (kb) between two consecutive loci, indicated that 81% of clones are derived from only 42 genomic regions (Figure 1.4). Moreover, 19% of the sequences in the 18–26–nt library also originate in these regions, indicating that the 26–nt shoulder in the size distribution profile of this library indeed corresponds to MILI–interacting RNAs. Even more surprisingly, we found that 76% of all sequences identified in the 24–33–nt library map to the same regions, which thus seem to produce both the 26–28–nt MILI–interacting RNAs and the abundant 29–31–nt RNAs that are present in total testis RNA. On the basis of the interaction of these small RNAs with the MILI member of the Piwi protein family,

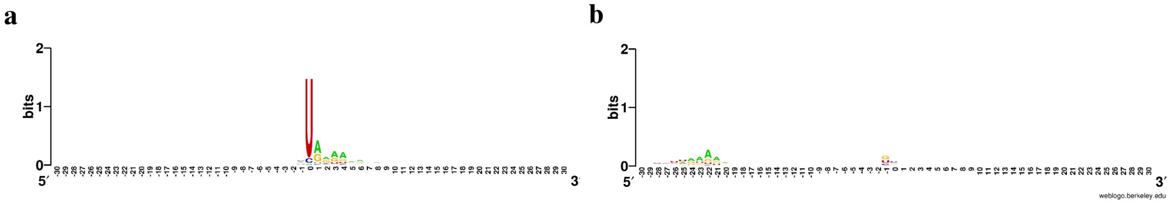


Figure 1.3: Graphic representation of an aligned set of the regions containing piRNAs. The genomic location of the clones with a unique mapping to piRNA clusters was used to extract 60 nt–long sequences, with 30 nt upstream and 30 nt downstream of the 5'–most nucleotide of the small RNA (located at position 0 in panel a) and of the 3'–most nucleotide of the small RNA (located at position 0 in panel b). The weblogo software (http://weblogo.berkeley.edu/logo.cgi) represents the positional weight matrix computed from the input sequences. The total height of the letters indicates the information score at that position, while the height of the individual letters indicates the relative frequency of each nucleotide at that position.

we refer to them as Piwi–interacting RNAs (piRNAs).

The pronounced clustering strongly suggests that multiple piRNAs are processed from long primary transcripts. A hypothesis supported by testis–specific expressed sequence tags (ESTs) and messenger RNAs mapping to these regions. The identified piRNA–encoding regions are distributed over most mouse chromosomes and they range in size from 0.9 to 127 kb (Figure 1.4). Although piRNAs map exclusively to one chromosomal strand in many regions, some regions–including the region on mouse chromosome 17 that is one of the major sources of piRNAs-encode piRNAs in both orientations. Notably, in all cases this arrangement was caused by two closely spaced clusters: one that almost exclusively contained sequences mapped to the sense strand and the other containing sequences mapped to the antisense strand, suggesting bidirectional divergent transcription from a central promoter (Figure 1.5). Thus, although piRNAs resemble in size the rasiRNAs of fruitfly and zebrafish [29, 31] they differ from rasiRNAs in several respects. First, piRNAs map to the genome in a highly strand–specific manner, in contrast to rasiRNAs that map to repeat regions in sense or antisense orientation as if randomly generated from long dsRNA precursors [29,31]. Second, piRNAs predominantly map to single genomic loci, whereas rasiRNAs map by definition to repetitive sites including transposable elements. In fact, the proportion

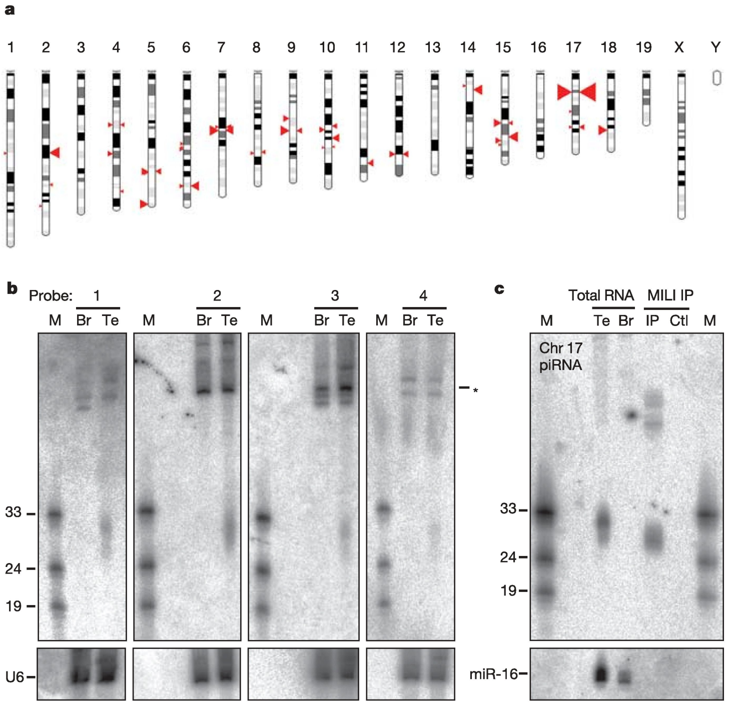


Figure 1.4: MILI–interacting piRNAs are encoded in clustered genomic loci. a, Location of piRNA–encoding regions on mouse chromosomes. The size of the triangles is proportional to the number of cloned MILI–interacting sequences. The left or right position of the triangles indicates mapping of the clones to the minus or plus strand, respectively. b, c, Northern blot hybridization for mouse chromosome (Chr) 17 piRNAs using partially hydrolysed RNA probes against four distinct 500–nt regions (b) or a 27–nt oligodeoxynucleotide probe complementary to a single piRNA clone (c) (top panels). Mouse total RNA from adult testis (Te) or brain (Br) was examined, and the size and mobility of the RNA size marker (M) is specified. The asterisk marks tRNA cross–hybridization signals. Blots were re–probed with a U6 antisense probe (bottom panels). MILI–immunoprecipitated RNAs (IP) are absent in control (Ctl) immunoprecipitation experiments; miR–16 is detectable in total RNA from testis or brain, but is undetectable in the MILI–immunoprecipitated sample.

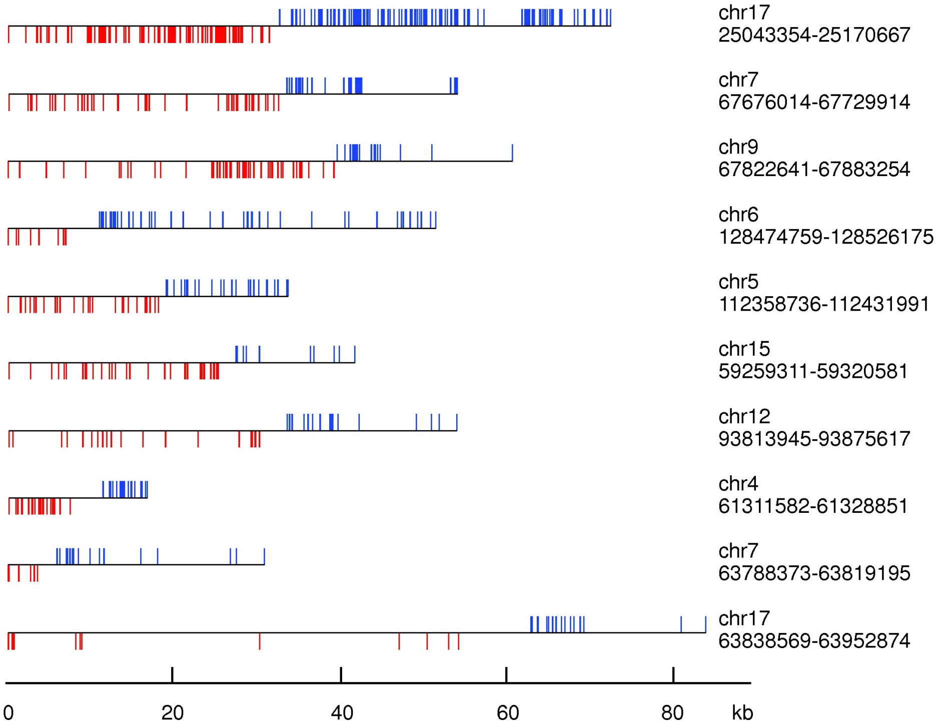


Figure 1.5: Bidirectional clusters of piRNA. Clones from the MILI–immunoprecipitated RNAs mapping to the genomic regions that define bidirectional transcript start sites are shown.

of repeat elements is smaller within the piRNA regions than in the 100–kb flanking regions (29% versus 38%, P–value , $2.2 * 10^{-16}$).

The spacing between cloned piRNAs within each genomic region has no apparent pattern. To assess whether the piRNAs show any evidence of specific processing, we aligned the piRNAs whose loci are partially overlapping (52% of the clones in the piRNA regions) and evaluated the precision with which their 5' and 3' ends have been processed. We found frequent examples of piRNAs whose 5' or 3' ends coincide, indicating that they are not random degradation products of long transcripts (Figure 1.6). Additionally, the 5' end seems to be more precisely processed than the 3' end, similar to what has been observed among miRNA clones [30]. To examine if piRNAs may be–like miRNAs–excised from dsRNA fold–back precursor structures, we investigated if any base–paired regions emerged in approximately 100 nt from each side of the piRNA. Although our computational approach was able to reveal the loop and stem regions of miRNA precursors, no clear hybridization pattern involving the piRNAs or the sequences flanking them was found (Figure 1.7). It is possible that long–range dsRNA structures or sequence–specific protein machinery are involved in guiding the maturation process. The strong preference for a 5' uridine in piRNAs suggests the involvement of Drosha or Dicer RNase III, but additional structural or sequence–specific determinants are yet to be identified.

The processing of piRNA primary transcripts was further examined by northern blotting of total testis RNA using probes specific to four 500–nt regions within the largest piRNA cluster on mouse chromosome 17. All four probes antisense to cloned piRNAs detected a signal at 26–31 nt in testis, but not in brain, RNA (Figure 1.4b). None of the sense probes yielded any signal, supporting the strand–specific accumulation of piRNAs (data not shown). Surprisingly, even a single oligodeoxynucleotide probe antisense to an individual piRNA was sufficient to detect a 26–28–nt–size signal in MILI–immunoprecipitated RNA and a 29–31–nt–size signal in total testis RNA (Figure 1.4c). Oligodeoxynucleotide probes to sequences immediately flanking the isolated piRNA failed to produce a signal (data not shown), suggesting that the processing of piRNAs occurs in a directed fashion. Examination of the 5' and 3' ends of a 30–nt cloned piRNA by rapid amplification of cDNA ends (RACE) showed that the 5' ends of the piRNA RACE clones were invariant in both libraries, whereas the 3' end clones were truncated by 2 nt in the MILI immunoprecipitate small–RNA library. It is possible that the more abundant 29–31–nt fraction represents an intermediate processing product or that it corresponds to small RNAs interacting with another Piwi protein expressed in testis. Re–probing of the RNA blot for the ubiquitously expressed 22–nt miR–16 produced a miRNA signal for testis total RNA but none for MILI–immunoprecipitated RNA, indicating that MILI was specifically loaded with piRNAs but not miRNAs.

To obtain insight into the temporal expression of piRNAs during mouse spermatogenesis, total testis RNA was then isolated at different time points of postnatal development. In mice, mitotically active spermatogonia represent the principal developing germ cells in testis up to day 6 after birth. Meiosis I is initiated on day 10, with germ cells reaching the preleptotene/leptotene, zygotene and early pachytene stages by days

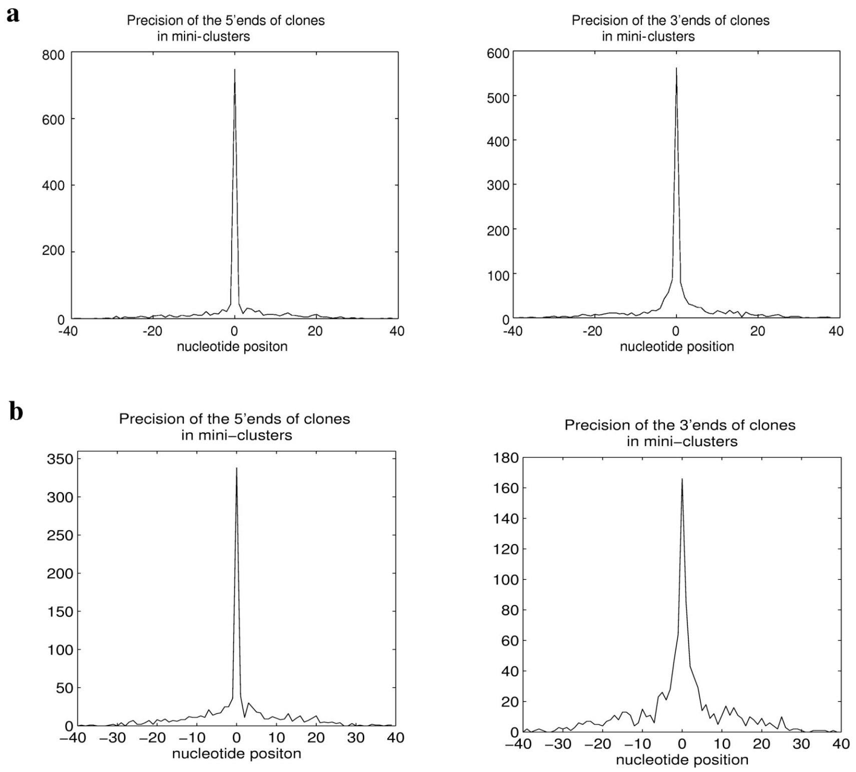


Figure 1.6: Precision of mouse piRNA processing at the 5' end (left panels) and the 3' end (right panels) for a) all single mapping clones and b) all single mapping unique clones (removing possible amplification biases resulting in multiple counts for the same small RNA). Partially overlapping clones from three libraries (52%) were aligned to form miniclusters. To assess the precision with which the 5' and the 3' ends of the piRNAs are processed, we determined the most frequently observed location of the 5' and 3' end, respectively, in each minicluster, and we constructed the histogram of the distances between the location of the 5' and 3' end of each sequence in the minicluster and the reference location of the 5' and 3' ends. The 5' ends of aligned sequences are more sharply defined than the 3' ends.

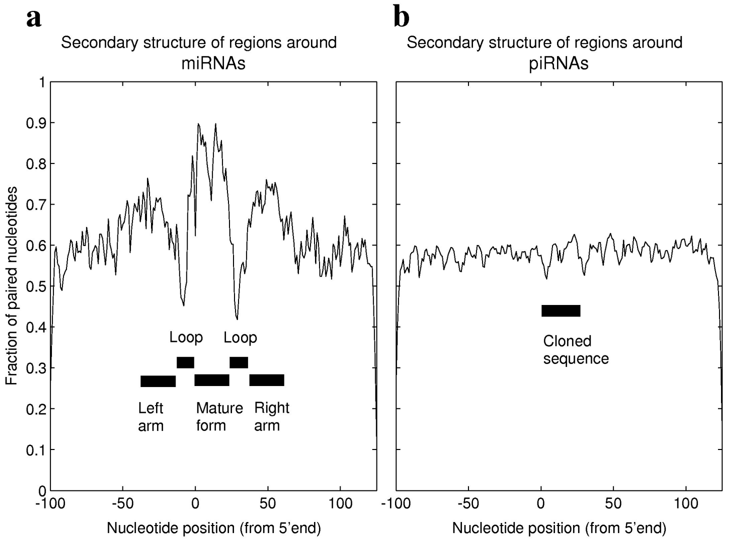


Figure 1.7: Propensity of regions around miRNAs and piRNAs to form secondary structures. The set of mouse miRNAs was extracted from the miRNA repository (http://microrna.sanger.ac.uk/sequences/index.shtml). The genomic location of the small RNA sequences (piRNAs or miRNAs) was used to extract 225 nt sequences, with 100 nt upstream and 125 nt downstream of the 5' end of the small RNA (located at position 0). These regions were folded using the RNAfold program of the Vienna package (http://www.tbi.univie.ac.at/ ivo/RNA), and the minimum free energy structure was used to determine an average profile of paired nucleotides along the sequence. The figure shows for each position the fraction of sequences whose nucleotide at that position was paired, over all miRNAs (a) and over all piRNAs (b). miRNAs clearly demonstrate secondary structure that involves mature miRNA and either left or right arm of the hairpin precursor (depending on whether the mature form of the miRNA is in the 5' or 3' arm of the pre–miRNA). No prominent secondary structure is seen for piRNAs.

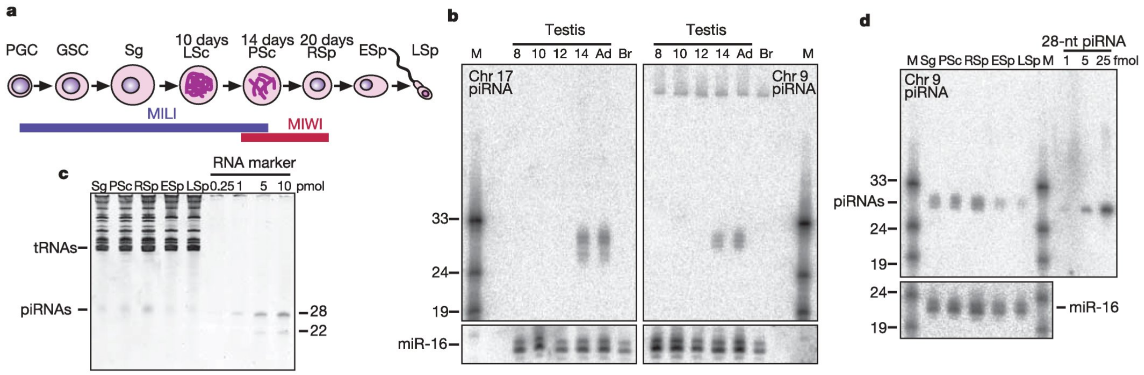


Figure 1.8: Temporal expression of piRNAs during mouse spermatogenesis. a, Sketch of mouse spermatogenesis with the temporal expression patterns of MILI (blue) and MIWI (red). Abbreviations: PGC, primordial germline cells; GSC, germline stem cells; Sg, spermatogonia; LSc, leptotene spermatocytes; PSc, pachytene spermatocytes; RSp, round spermatids; ESp, elongating spermatids; LSp, late spermatids. b, Northern blot of testis total RNA from 8–, 10–, 12– and 14–day–old newborn and 3–month–old adult (Ad) mice and brain (Br) total RNA with oligodeoxynucleotide probes complementary to piRNAs (top panels) from chromosome (Chr) 17 (left panel) and chromosome 9 (right panel). The asterisk marks a cross–hybridization signal to a larger RNA also present in brain. Expression of miR–16 is monitored and serves as loading control (bottom panels). c, SYBR Green II staining of total RNA from elutriator–enriched male germ cells separated on a denaturing polyacrylamide gel. As reference RNA markers, two 22–nt and two 28–nt oligoribonucleotides of distinct sequences were loaded. d, Northern blot for samples shown in c using antisense probes specific for the piRNA cluster on chromosome 9 (top panel) and for miR–16 (bottom panel). A synthetic 28–nt chromosome 9 piRNA was loaded to quantify the amount of piRNA expressed in germ cells.

10, 12 and 14, respectively [32] (Figure 1.8a). MILI is expressed in male germ cells from primordial germ cells until the pachytene stage of meiosis [27], whereas MIWI is expressed in pachytene–stage spermatocytes and round spermatids [28]. Northern blotting of testis total RNA for two distinct piRNAs from mouse chromosome 9 and 17 revealed piRNA accumulation starting at day 14, when the first spermatocytes reach the pachytene stage (Figure 1.8b). To assess the presence of piRNAs in specific germ–cell types, RNA was isolated from cells enriched for different stages of spermatogenesis after elutriation purification.Notably, the 30–nt piRNAs were so abundant that they were visible by SYBR Green II staining, and were estimated to be present at about 1 million piRNA molecules per mouse spermatocyte or round spermatid (Figure 1.8c). Quantitative northern blotting for an individual piRNA from mouse chromosome 9 indicated about 8,000 copies per pachytene spermatocyte and about 2,000 copies per haploid round spermatid (Figure 1.8d). The piRNA level was reduced by about tenfold in RNA isolated from later stages of germ–cell development.

Mouse piRNA sequences are well conserved and cluster within the closely related rat genome. However, the alignments of the mouse genome with eight other genomes [33] indicated that piRNA regions are poorly conserved between more distant species (Figure 1.9), and that conserved elements are present with similar frequency within piRNA clusters and within introns of proteincoding genes (Figure 1.9b). For seven of the ten mouse piRNA clusters that seem to contain a bidirectional promoter, we found short regions of homology with the human genome. Moreover, we found that the frequency of ESTs that overlap with these human genomic regions orthologous to the mouse piRNA clusters was 9–21–times higher compared with the representation of testis ESTs among all the GenBank ESTs. To provide experimental support for human piRNAs, 18–26–nt and 24–33–nt small–RNA libraries from human testis total RNA (Figure 1.2c) were further prepared and sequenced. The small– RNA composition shows the expected enrichment for 5' uridine, especially in the longer–size library where a 5' uridine bias is not introduced by the presence of miRNAs. Using the same clustering criteria as for the mouse sequences, we were able to define 14 human piRNA clusters. They, together, contain 8.5% of all the cloned human sequences, and 24% of the human clones that were not derived from other functional RNAs. The divergently transcribed piRNA cluster with the strongest expression in

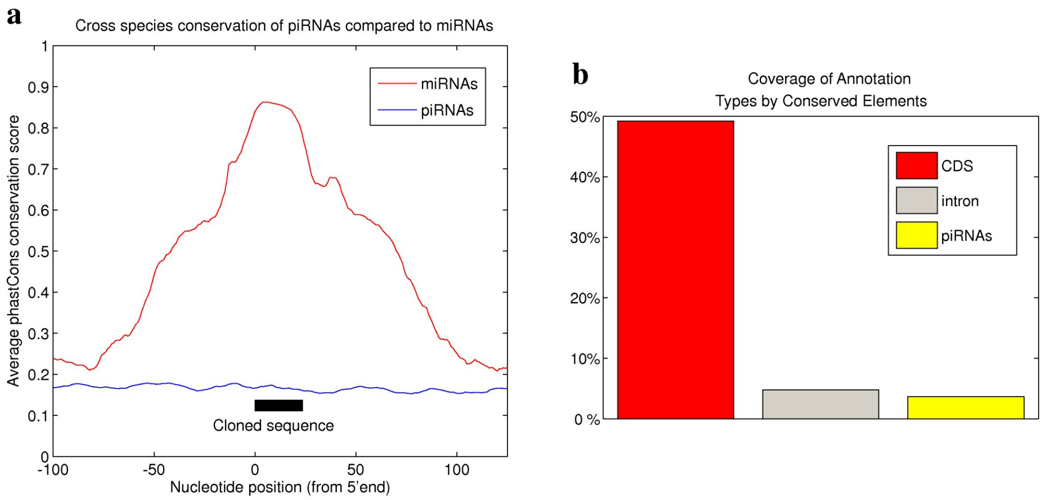


Figure 1.9: Cross–species conservation of the individual piRNAs and piRNA clusters. a, Cross–species conservation profile of piRNAs compared with miRNAs. The phastCons conservation scores from the UCSC database were used as a measure of nucleotide conservation (see Methods). The plot shows the conservation score at each position, averaged first over the miRNAs (red) and then over the piRNAs (blue). b, Cross–species conservation of the piRNA regions. The phastCons conserved elements that overlap piRNA regions were extracted from the UCSC database. The coverage of piRNA regions by conserved elements in comparison with CDS and intronic regions of mouse RefSeq mRNAs [34], was determined similarly to a previously described analysis [33].

mouse is orthologous to the divergently transcribed cluster with the strongest expression in human (Figure 1.10), and two additional human regions that are orthologous to divergently transcribed mouse piRNA clusters are experimentally supported.

Although Piwi proteins were shown to be important for stem– and germ–cell development in different animals [25–28], the underlying biochemical pathways are unknown. The identification of piRNAs provides an important molecular link regarding the function of Piwi proteins. Given the timing of the maturational arrest in mili and miwi knockout mice at the pachytene spermatocyte [27] and the spermatid steps [27], respectively, it is conceivable that piRNAs and germline–specific Piwi proteins regulate the timing of meiotic and postmeiotic events through transcriptional and translational repression. Argonaute proteins have been implicated in diverse

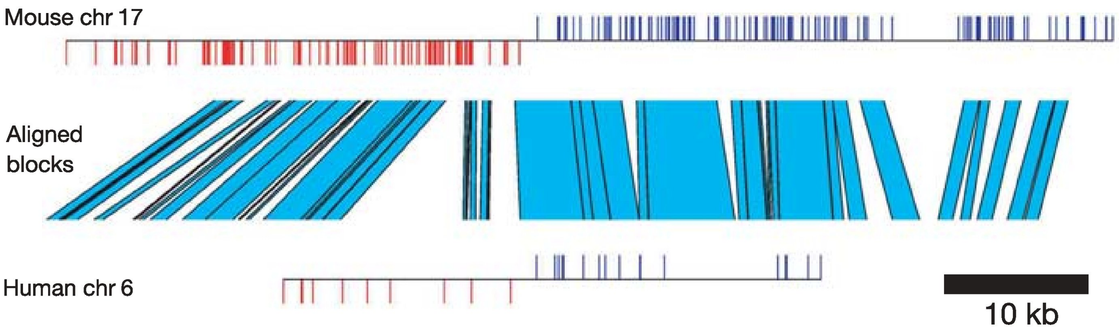


Figure 1.10: Predominant mouse piRNA cluster and its orthologous cluster in human. Alignment view of the most highly expressed mouse piRNA cluster and its corresponding human orthologue. The positions of cloned sequences indicate divergent transcription from a central promoter in both species.

processes such as genome rearrangement in Tetrahymena [35] or heterochromatic silencing and chromosome segregation in fission yeast [36], and we are only beginning to develop an understanding of the molecular mechanisms mediated by the diverse group of Argonaute ribonucleoprotein complexes.

1.3.4 Methods

Preparation of male germ cells and testis extracts. Germ cells were obtained from the seminiferous tubules of 3–month–old C57BL/6J male mice (Jackson Laboratory, Bar Harbor, ME) by the separation and purification of spermatogenic cells on the basis of sedimentation velocity using centrifugal elutriation as previously described [37]. Pachytene spermatocytes ($2.3 * 10^7$ cells) yielded 420 μg and round spermatids ($9.8 * 10^7$ cells) 270 μg of total RNA. Twenty–four testicles were washed with ice–cold PBS and homogenized in two volumes of buffer (25 mM Tris–HCl, pH 7.5, 150 mM KCl, 2 mM EDTA, 0.5% NP40, 1 mM NaF, 1mM DTT, 100 U/ml RNasin ribonuclease inhibitor (Promega), Complete EDTA–free protease inhibitor (Roche) with a Dounce homogenizer. The concentrated testis lysate was cleared by centrifugation in a Sorvall fresco tabletop centrifuge at 14,000 rpm (16,000 g) for 10 min at 4°C. The total protein concentration of the extract was about 35 mg/ml.

Immunoprecipitation of MILI ribonucleoprotein complexes, isolation and labelling of bead–bound nucleic acids. For immunoprecipitation, 1.2 ml of cleared lysate was diluted 12.5 fold to a final protein concentration of 2.8 mg/ml with NT2 buffer (50 mM Tris–HCl, pH 7.4, 150 mM NaCl, 1 mM $MgCl_2$, 0.05% NP40) supplemented with 1 mM DTT, 2 mM EDTA and 100 U/ml RNasin. Protein A Sepharose CL–4B beads (150 μl, Sigma, P3391) were equilibrated with NT2 buffer and incubated with 15 μl of 1.7 mg/ml affinity–purified anti–MILI–pepN2 antibody raised against the peptide VRKDREEPRSSLPDPS (amino acids 107-122) for 6 hours at 4°C with gentle agitation. The diluted testis lysate was added to the beads and the incubation was continued for overnight at 4°C. The beads were washed twice with ice–cold NT2 and twice with NT2 with the concentration of NaCl adjusted to 300 mM. Control immunoprecipitations were carried out in the absence of the antibody. Nucleic acids that co–immunoprecipitated with MILI were isolated by treatment of the beads with 0.6 mg/ml proteinase K in 0.3 ml proteinase K buffer, followed by phenol (at neutral pH)/chloroform extraction and ethanol precipitation. Nucleic acids that co–immunoprecipitated with MILI were isolated by treatment of the beads with 0.6 mg/ml proteinase K in 0.3 ml proteinase K buffer, followed by phenol (at neutral pH)/chloroform extraction and ethanol precipitation. For 5' labelling, aliquots of the isolated nucleic acids were first subjected to dephosphorylation with calf intestinal phosphatase as described [14]. After phenol/chloroform extraction and ethanol precipitation, the RNAs were labelled with $[\gamma -^{32} P]$-ATP by T4 polynucleotide kinase and resolved on a 15% acrylamide gel along with radioactive oligoribonucleotide size markers.

Cloning of small RNAs. Total RNA from mouse testis was prepared as previously described [38]. A previously prepared size–fractionated testis library of 18– to 26–nt RNAs [38] was re–amplified and subjected to large–scale sequencing. A new small RNA library covering the size range of 24– to 33–nt was prepared using pre–adenylated 3' adapters as described [39]. The same revised protocol was used to clone MILI–associated small RNAs without size selection, but by adding a trace amount of 5'–labelled immunoprecipitated small RNA described above. Human total RNA used for the preparation of the 18– to 26–nt and 24– to 33–nt library was purchased from Ambion (22–year old male), or prepared by M. J. Brownstein from testis of a 73–year old male.

Northern blot analysis and piRNA quantification. Northern blots for detection of miRNAs and individual piRNA were performed, as described previously loading 10 μg of total RNA per well [38]. The oligodeoxynucleotide probes for piRNAs on chr.9 and 17 were 5' TCCCTAGGAGAAAATACTAGACCTAGAA and 5' TCCTTGTTAGTTCTCACTCGTCTTTTA, respectively, and for miR–16 and U6 snRNA 5' GCCAATATTTACGTGCTGCTA and 5' GCAGGGGCCATGCTAATCTTCTCTGTATCG, respectively. The content of chr.9 piRNA in male germ cells was determined by quantitative Northern blotting using synthetic 5' UUCUAGGUCUAGUAUUUUCUCCUAGGGA for calibration. To quantify total piRNAs in germ cells by SYBR Green II staining, 10 μg of total RNA were loaded per well. The 22– and 28–nt reference standard contained equimolar amounts of 5' AACUGUGUCUUUUCUGAAUAGA and 5' UAUUUAGAAUGGCGCUGAUCUG or 5' UAAAAGACGAGUGAGAACUAACAAGGAG and 5' UUCUAGGUCUAGUAUUUUCUCCUAGGGA, respectively. SYBR Green staining is sequence dependent so that the 22–nt and the 28–nt reference standards yield somewhat different fluorescence intensities. The RNA probes that cover fragments of piRNA–containing regions were produced from about 500–nt long internally [$\alpha -^{32}P$]-UTP-labelled T3 or T7 RNA polymerase in vitro transcripts using PCR templates amplified from mouse genomic DNA by three rounds of nested PCR. The transcripts were partially hydrolysed in the presence of one volume of carbonate buffer (60 mM Na_2CO_3, 40 mM $NaHCO_3$) at 60°C for 7 min. Time of hydrolysis was chosen in pilot experiments to generate fragments with length of 50– to 100–nt. After neutralization with 200 mM HCl, probes were further purified by gel filtration through G–25 columns (Amersham). The hybridization using these probes was performed at 50°C in 5x SSC, 20 mM Na_2HPO_2, pH 7.2, 7% SDS, 1x Denhardt's solution, 30% (v/v) formamide. The membrane was washed twice with 2x SSC, 1% SDS solution and twice with 0.5x SSC 1% SDS at 50°C.

RACE. For 5' RACE, 2 μl of the mixture of reverse transcription reaction from the small RNA cloning step was amplified with a universal forward primer that matches the 5' adapter sequence and reverse primer to chr. 17 piRNA (5' TCCTTGTTAGTTCTCACTC). For 3' RACE, a specific sense primer (5' TAAAAGACGAGTGAGAACTA) and a universal reverse primer to the 3' adapter were used. The primers shown above were labelled by T4 polynucleotide kinase with [$\gamma -^{32}P$]-ATP and added to the PCR reaction at 0.06 μM final concentration together with 0.5 μM of forward and reverse non–labelled primers. 25 cycles of PCR amplification were performed at 94°C for 50 s, 50°C for 40 s and 72°C for 30 s. PCR products were mixed with formamide loading buffer, denatured briefly at 90°C and resolved on 8% polyacrylamide gel and the resolved bands were examined by phosphorimaging. For cloning and sequencing, RACE PCR products prepared with unlabelled primers were ligated into pCR2.1–TOPO (Invitrogen).

Genome mapping and functional annotation of cloned small RNA. Cloned small RNAs were mapped to the mm6 assembly of the mouse genome and to sequences with known function, to infer the likely origin of the cloned RNAs. The genome assembly and some functional annotation are available from the genome browser at the UCSC (http://genome.ucsc.edu). The mappings were performed using the Washington University implementation (http://blast.wustl.edu, W. Gish, 1996–2004) of BLAST as well as in–house sequence alignment programs. For each small RNA sequence we only used the best matches up to maximum three differences (mismatch, insertion or deletion) for subsequent analyses. The functional annotation was done as described before [13,31,39]. The database of sequences with known function was assembled from rRNA, tRNA, snRNA, snoRNA, scRNA (small cytoplasmic RNA) and mRNA sequences obtained by querying GenBank (http://www.ncbi.nih.gov/Genbank/index.html), with the appropriate feature key. We additionally used a data set of non–coding RNAs from the NONCODE database (http://noncode.bioinfo.org.cn), the miRBase database of miRNAs (ftp://ftp.sanger.ac.uk/pub/mirbase/sequences/CURRENT/), the snoRNA database (http://www-snorna.biotoul.fr), predicted miRNA sequences [40–42]. For the repeat annotation, we used the repeat masker results from the UCSC database. To count the number of sequences derived from a particular class of repeats, we intersected the genomic loci of the clones with the genomic regions that were annotated with that class of repeats. The genomic locus was considered to be repeat–associated if it overlapped by at least 15 nucleotides with an annotated repeat element. Sequences that mapped to piRNA clusters (defined below), and did not match other known functional RNAs or repeat elements were called piRNAs.

Definition of piRNA clusters. piRNA clusters for mouse were defined using the following criterion: two genomic loci corresponding to small RNAs cloned from the

MILI IP library were placed in the same cluster if they were less than 15 kb apart in the genome, irrespective of their strand. Once the cluster boundaries were identified this way, we determined the number of small RNAs that originated in each cluster, and retained only those regions with at least 4 sequences. Given that some small RNAs map to multiple locations in the genome, we assumed that each of these locations is equally likely to have produced the small RNA. Therefore, the number of sequences originating in each of these locations was defined as the number of times the sequence was cloned divided by the number of genomic loci in which the sequence could have originated. For human piRNAs, the 24– to 33–nt library was used to define initial piRNA clusters. We first eliminated the sequences derived from rRNA, tRNA, snRNA, snoRNA and miRNAs, and then we clustered the remaining sequences as we did for mouse.

Coverage of piRNA clusters by repeat elements. To reveal the fraction of piRNA regions covered by repeat elements we used the repeat masker results from the UCSC database to determine the proportion of nucleotides within the piRNA clusters and within 200 kb (100 kb on each side) around the piRNA regions that are covered by repeat elements. 450141 of the total 1534522 nucleotides in the piRNA regions (29.3%) and 3016211 of the 7992650 (37.7%) in the flanking regions overlapped with annotated repeat elements.

Precision of mouse piRNA processing at the 5' end and the 3' end. Partially overlapping clones from three libraries (52%) were aligned to form miniclusters. We then determined the most frequently observed location of the 5' and 3' end, respectively, in each minicluster, and we constructed the histogram of the distances between the location of the 5' and 3' end of each sequence in the minicluster (not including the reference sequence) and the reference location of the 5' and 3' ends. We verified that our results hold even when we use only one copy of each sequence that was cloned multiple times within a give library, thus excluding the possible effects of multiple amplification products of the same RNA within a library.

Propensity of regions around miRNAs and piRNAs to form secondary structures. The set of mouse miRNAs was extracted from http://microrna.sanger.ac.uk/. The genomic location of the small RNA sequences (piRNAs or miRNAs) was used to extract 225 nt sequences, with 100 nt upstream and 125 nt downstream of the 5' end of the small RNA (located at position 0). These regions were folded using the RNAfold program of the Vienna package (http://www.tbi.univie.ac.at/ ivo/RNA), and the minimum free energy structure was used to determine an average profile of paired nucleotides along the sequence.

Cross–species conservation of the individual piRNAs and of the piRNA clusters. The genomic mapping of the small RNA sequences (piRNAs or miRNAs) was used to extract 225 nt sequences, with 100 nucleotides upstream and 125 downstream of the 5' nucleotide of the small RNA (located at position 0). The phastCons [33] conservation scores were obtained from the UCSC annotation of the mm6 assembly version of the mouse genome (http://hgdownload.cse.ucsc.edu/downloads.html#mouse). We then computed the average phastCons score at every position in the regions around miRNAs and piRNAs. We additionally obtained the phastCons [33] conserved elements from the same source, and we extracted those that overlap piRNA regions. We then determined the coverage of piRNA regions by conserved elements and compared it with the coverage of CDS and intronic regions of mouse RefSeq mRNAs [34], computed as described in a previous analysis [33]. To determine the human orthologs of mouse piRNA clusters we used the following procedure. We focused on the mouse piRNA–encoding regions that contained the putative bidirectional promoters, because for these, the mapping of the cloned sequences gives us a good indication of the location of the promoter. From the whole genome alignments provided on the Genome Bioinformatics Site at UCSC we selected for each of the mouse promoters the largest alignment block that overlaps with it, and we used this as an anchor in the orthologous region of the human genome. We were only able to extract human anchors for 7 of the 10 mouse promoter regions. We then selected the regions extending 30 kb on each side of each of the human anchors to identify ESTs that overlap with, and were therefore expressed from the human regions that are orthologous to the mouse piRNA bidirectional clusters. We used the Genbank records for these ESTs to identify those that appear to have been isolated from testis (based on the clone_lib or tissue_type fields of the Genbank record). For comparison, we determined the proportion of testis–expressed ESTs among all the ESTs that have been mapped to the human genome by the UCSC Genome Bioinformatics Group.

1.3.5 Acknowledgements

Grant support. This work was supported by a FRAXA Research Foundation postdoctoral fellowship to A.A., an NIH grant HD39024 to P.M., NIH grants R01 GM068476-01 and P01 GM073047-01 to T.T., and an SNF grant 205321-105945 to M.Z.

Table 1.1: Characterization of MILI IP and testis total small RNA libraries. Small RNA clones sequenced from MILI IP and testis total RNA libraries were mapped to the mouse or human genomes and annotated as described in Methods. The 18– to 26–nt library from mouse displays a high rRNA content because its library preparation protocol, in contrast to other libraries listed, did not require a 5' phosphate on the isolated RNAs to be represented in the library. [1]Unique clones indicate the fraction of sequences that were cloned only once in a given library. [2]Two sequences were clustered together if they mapped closer than 15 kb from each other. Clustering was done independently for the five libraries. We selected only clusters containing at least 4 sequences.

Features		Mouse testis total RNA libraries		MILI IP	Human testis total RNA libraries	
		18- to 26-nt	24- to 33-nt		18- to 26-nt	24- to 33-nt
Number of clones		13312	805	1673	2054	619
Average size ± st. dev. (nt)		21.89 ± 3.18	29.47 ± 1.97	26.88 ± 2.28	21.89 ± 2.73	28.57 ± 2.75
Unique clones[1] (%)		79.00	92.90	97.73	81.46	90.97
Uridine in 5' position (%)		48.11	88.45	84.52	64.22	59.77
Clustered within 15 kb[2] (%)		80.04	78.42	81.15	69.33	47.46
Fraction of small RNA clones (in %) that match to the genome with 0 to >10 times, as indicated.	0	0.67	2.86	2.81	0.83	3.23
	1	43.28	88.82	87.57	68.26	50.57
	2-3	18.37	3.98	5.68	21.28	10.82
	4-10	28.74	3.11	1.79	4.82	18.26
	>10	8.94	1.24	2.15	4.82	17.12
Annotation (% clones)						
rRNA		34.44	1.49	0.54	5.55	2.43
tRNA		1.97	1.37	0.78	1.46	23.62
miRNA		22.90	0.25	0.30	67.09	0.97
sn/snoRNA		1.90	0.37	0.12	1.07	1.62
piRNA		16.59	67.20	72.62	2.34	25.73
mRNA		10.12	6.09	5.74	9.06	16.67
repeat sequence		7.69	13.91	12.79	7.06	14.89
none		4.11	9.19	7.05	6.09	14.08

Table 1.2: The sequences that match to sense (+), antisense (-) of each repeat type are indicated. For each repeat type, we considered all genome locations, which were annotated with that repeat type in the UCSC database. Since a repeat–annotated sequence maps to multiple locations in the genome, each of these loci was considered to have potentially given rise to the sequence with probability 1/number of loci. We then summed over all loci of a given type the probabilities of each sequence arising from all these loci. Some genomic regions have multiple repeat annotations. Each of these was considered separately, and thus the number of counted repeats is somewhat larger than the number of sequences cloned from repeats.

Repeat type	Number of clones in MILI IP library	Strand orientation (+)/(-)
DNA	13	7(+)/6(-)
MER2_type	2	2(+)
MuDR	1	1(+)
AcHobo	1	1(+)
MER1_type	9	3(+)/6(-)
LINE	46	27(+)/19(-)
L2	9	5(+)/4(-)
RTE	1	1(+)
L1	36	21(+)/15(-)
LTR	117	38(+)/79(-)
ERVL	17	6(+)/11(-)
MaLR	37	6(+)/31(-)
ERVK	44	10(+)/34(-)
ERV1	19	16(+)/3(-)
SINE	65	37(+)/28(-)
Alu	23	8(+)/15(-)
B4	15	13(+)/2(-)
B2	9	6(+)/3(-)
MIR	18	10(+)/8(-)
Satellite	1	1(+)
Satellite-unspecified	1	1(+)
Other	1	1(-)
Other-unspecified	1	1(-)
Total	243	243

Chapter 2

Oligomap: a program for fast identification of nearly-perfect matches of small RNAs in sequence databases

Parts of this section will appear in a special issue on miRNAs of Methods in Enzymology 2008.

2.1 Introduction

A keys step in the process of small RNA annotation requires the small RNAs to be mapped to sequences of known function and to the corresponding genome. This process has to be sensitive, meaning that all small RNAs that do have matches within the specified quality constraints should be mapped, and efficient, meaning that the program should not take longer than a day to map millions of small RNAs. Variants of the Blast algorithm [43], such as WU-BLAST (http://blast.wustl.edu) [15], Blast [44, 45], or Megablast ([46]) [47] have been used for this purpose as well. Typically, the output of these programs is filtered to retain only very good alignments, with very few differences between small RNAs and targets. The programs mentioned above are in fact very general, but they have been designed for mapping longer RNAs (such as ESTs), and in order to achieve good performance, they use heuristics, such as initiating alignments from perfect contiguous matches of a minimum length (”words”) between query and target sequence. Because sequencing errors in 18-30-nucleotides long RNAs can easily reduce the length of the contiguous matches to the target sequence, one would have to use a relatively small word size in order to guarantee that 1-error hits are retrieved, thereby increasing the running time of the programs. This becomes a problem when we need to map hundreds of thousands of small RNAs to mammalian genomes. Moreover, if all we want to do in the end is to identify very close matches of short RNA sequences, the complexity of these general algorithms is not necessary. We therefore developed a special-purpose mapping algorithm that allows us to rapidly and *exhaustively* identify all the perfect and 1-error (where an error is defined to be a mismatch, insertion or deletion) matches of large sets of small RNAs to target sequences. The program can be downloaded from http://www.mirz.unibas.ch/software/.

2.2 Oligomap algorithm

A sketch of the main components of the algorithm is shown in Figure 2.1. The approach is to build a tree from the input small RNA sequences (Figure 2.1C) and then search this tree with subsequences starting at each position of the target sequence (Figure 2.1D). Each node in the tree corresponds to a nucleotide, and each small RNA is represented in the tree as a path that starts at the root and ends at either another internal node or at a leaf. There are 4 possible links from a parent node to a child node, one corresponding to each of the nucleotides. The identifier (ID) of each node encodes information about the small RNA represented by the path starting at the root and ending at the respective node (Figure 2.1A). The search stage is performed through a number of ”walkers” (Figure 2.1B). A walker represents a suffix of the target sequence that ends at the current position in the target. Every time a walker visits a node that represents a small RNA, we report a match between that small RNA and the target. When a walker ends in an internal node that does not represent a small RNA, it is removed from the search.

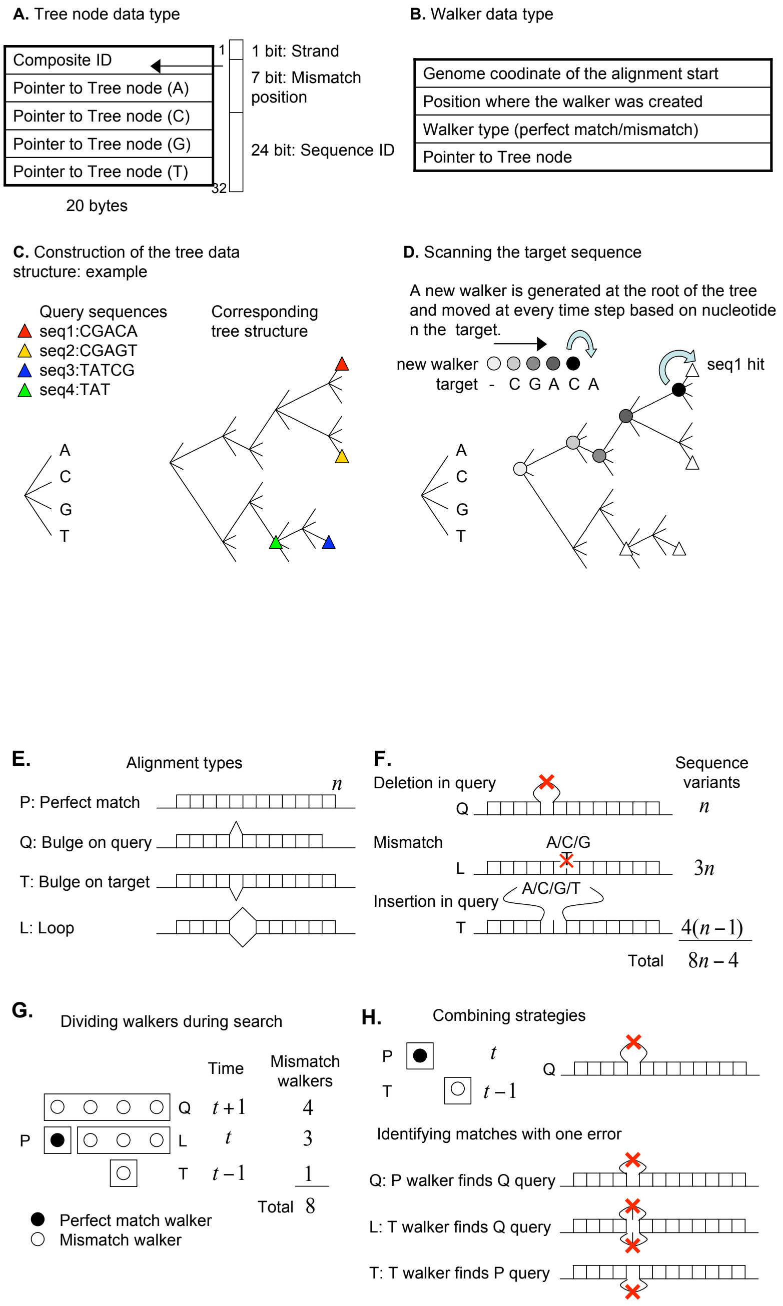


Figure 2.1: Schematic representation of the mapping algorithm.

2.2.1 Estimation of the resource requirements

To gain insight into the resource requirements of our algorithm, it is instructive to first consider the simple case in which we only want to identify perfect matches between small RNAs and a target sequence. For simplicity, let us assume that all small RNAs have the same length L. Then every small RNA will be represented as a path from root to a leaf in the tree and to construct the tree from N input sequences we need to visit $N * L$ nodes. Thus, the time needed for constructing the tree is proportional to $N * L$. The search phase consists of following paths in this tree starting from every nucleotide in the target. To do this, we start at the root of tree and visit the child which corresponds to the nucleotide currently observed in the target. We then continue on this path using the next nucleotide in the target and so on, until we either reach a leaf, or until the the internal node does not have a child that corresponds to the current nucleotide in the target. The length of a path that starts at a given nucleotide in the target determines the time needed to decide whether this path specifies an input small RNA. With L being the length of a small RNA, the upper bound on the path length is L, which for our applications is $20 - 35$. The average path length that we more typically encounter is however much shorter, as shown by the following argument. Assume that we generate the tree from N random sequences of length L defined over an alphabet of size A. Then the average length of a path that we will traverse starting from a given position in the target is given by the sum is over all possible path lengths l, the length of the path multiplied by the probability that the search will stop *precisely* after l steps. This will happen when none of the N sequences inserted in the tree had the prefix of length $l + 1$ of the sequence that we are searching for, but did have the prefix of length l. The average number of steps is thus given by:

$$
\begin{aligned}
S &= L\left(1-\left(1-\frac{1}{A^L}\right)^N\right) \\
&+ \sum_{l=1}^{L-1} l\left[\left(1-\left(1-\frac{1}{A^l}\right)^N\right)-\left(1-\left(1-\frac{1}{A^{l+1}}^N\right)\right)\right] \\
&= L\left(1-\left(1-\frac{1}{A^L}\right)^N\right)+\sum_{l=1}^{L-1} l\left(\left(1-\frac{1}{A^{l+1}}^N\right)-\left(1-\frac{1}{A^l}\right)^N\right)
\end{aligned}
$$

$$= \quad L - \sum_{l=1}^{L} (1 - \frac{1}{A^l})^N. \tag{2.1}$$

As shown in Figure 2.2, this number grows approximately logarithmically with N. For the values of A, L and N that are typical for our applications $(4, 22, 500000$, respectively), the average path will be approximately 9. The search time thus depends linearly on the target size and approximately logarithmically on the number of small RNAs.

The memory requirements of this program are determined by the size of the tree that we construct from the input small RNAs, an upper bound on this being $k*N*L$, with k a constant. An average estimate of the memory requirements can be obtained as follows. Given a tree in which $n-1$ sequences were already inserted, we want to compute the number of new nodes that the insertion of the n^{th} sequence will create. When processing the n^{th} sequence, a new node will be generated at level l in the tree if none of the sequences observed up to that point had the same length l prefix as sequence n. This happens with probability

$$\left(1 - \frac{1}{A^l}\right)^{n-1}.$$

Thus, inserting the n^{th} sequence will result, on average, in the insertion of

$$m(n) = \sum_{l=1}^{L} \left(1 - \frac{1}{A^l}\right)^{n-1}$$

nodes. Inserting progressively a total of N sequences generates on average

$$M(N) = \sum_{n=1}^{N} m(n) = \sum_{n=1}^{N} \sum_{l=1}^{L} \left(1 - \frac{1}{A^l}\right)^{n-1}. \tag{2.2}$$

Exchanging the two summations and applying the geometric series formula we obtain

$$M(N) = \sum_{l=1}^{L} \frac{1 - \left(1 - \frac{1}{A^l}\right)^N}{1 - \left(1 - \frac{1}{A^l}\right)}. \tag{2.3}$$

Finding 1-error matches requires that we either enumerate all these variants of the input small RNAs and insert them in the tree, or that we search the tree in such a way that we can identify matches with 0 or 1 error. The first option requires considerable more memory, since for every small RNA of length L we will have $8*L-4$

variants with 1 error (see Figure 2.1F). The search time would increase comparatively little, because the path length increases very slowly with the number of small RNAs represented in the tree. On the other hand, the second option requires little extra memory, but has a considerably longer search time, since at each position in the target we need to search not only for a perfect match starting at that position, but also for all the possible matches with 1 error (Figure 2.1G). This means following 8 additional search paths from each node on the path representing a perfect match of the target to a small RNA.

To achieve a good tradeoff between memory and CPU usage, we have combined these two strategies (Figure 2.1H): we store in the tree only the small RNAs (which we call P small RNAs) and their 1-nucleotide *deletion* variants (which we call Q small RNAs). Then, in the search process we create walkers representing target subsequences (P walkers) and their 1-nucleotide *deletion* variants (which we call T walkers). The 0− and 1-error variants of the small RNAs will be detected as follows:

1. perfect match small RNA-target: P walker stops at P small RNA
2. deletion in small RNA: P walker stops at Q small RNA
3. deletion in target: T walker stops at P small RNA
4. mismatch small RNA-target: T walker stops at Q small RNA, and looped out nucleotides do not match

Using the same argument that we used above, we can compute the average number of steps required to decide whether a path that starts at a given nucleotide in the target specifies an input small RNA. The difference is that the hybrid algorithm does not use a single walker starting from a given nucleotide in the target, but it spawns new ones from every point along the path of a perfect walker. The probability that these stop at a particular level l is the same as for a perfect walker, but the number of steps that they perform is smaller: if a T walker started at level h, it will only perform $l - h + 1$ steps up to level l. Thus, the average total number of steps performed by the P and T walkers initiated from a given position in the target is given by

$$
\begin{aligned}
S &= \sum_{h=1}^{L}\left[(L-h+1)\left(1-\left(1-\frac{1}{A^{L-h+1}}\right)^{N}\right)\right] \\
&+ \sum_{h=1}^{L}\left[\sum_{l=h}^{L-1}(l-h+1)\left(\left(1-\frac{1}{A^{l+1}}\right)^{N}-\left(1-\frac{1}{A^{l}}\right)^{N}\right)\right] \\
&= \frac{L(L+1)}{2}-\sum_{l=1}^{L} l\left(1-\frac{1}{A^{l}}\right)^{N}. \qquad (2.4)
\end{aligned}
$$

The behavior of these functions of N are shown in Figure 2.2A for $A = 4$ and $L = 16, 20, 24, 28, 32, 36, 40$.

2.2.2 Algorithm performance in a realistic setting

To illustrate the performance of our program particularly on very large sequence datasets for which it was designed, we used instead of small RNAs, for which large-scale data sets are only starting to be generated, the CAGE tag data generated by the Riken Institute in Japan [48]. These are short (20-21 nucleotides) sequences from the 5' ends of capped mRNAs, and millions of such sequences are already available. We constructed from this dataset 5 random subsets of sizes from $1,000$ to $512,000$ sequences, which we then mapped to the mouse genome assembly using our program. Figure 2.2C shows that the running time of the program increases only by a factor of 10 as the number of sequences in the input increases by a factor of 512. Mapping half a million sequences to the entire mouse genome takes roughly 5 hours on a 2.2 GHz AMD Opteron, using 2.3 GB of memory. We use this program to identify all close matches of small RNAs to their corresponding genome, and to other RNAs whose function is already known. The program can be downloaded from http://www.mirz.unibas.ch/software/ .

2.2.3 Implementation of oligomap

Oligomap is implemented in C++. The code is very compact and the program uses no external library. Since oligomap indexes the query sequences, there is no need to create chromosome index files. Both input files (targets and queries) have to be provided in fasta format. Running oligomap with no parameters will return an

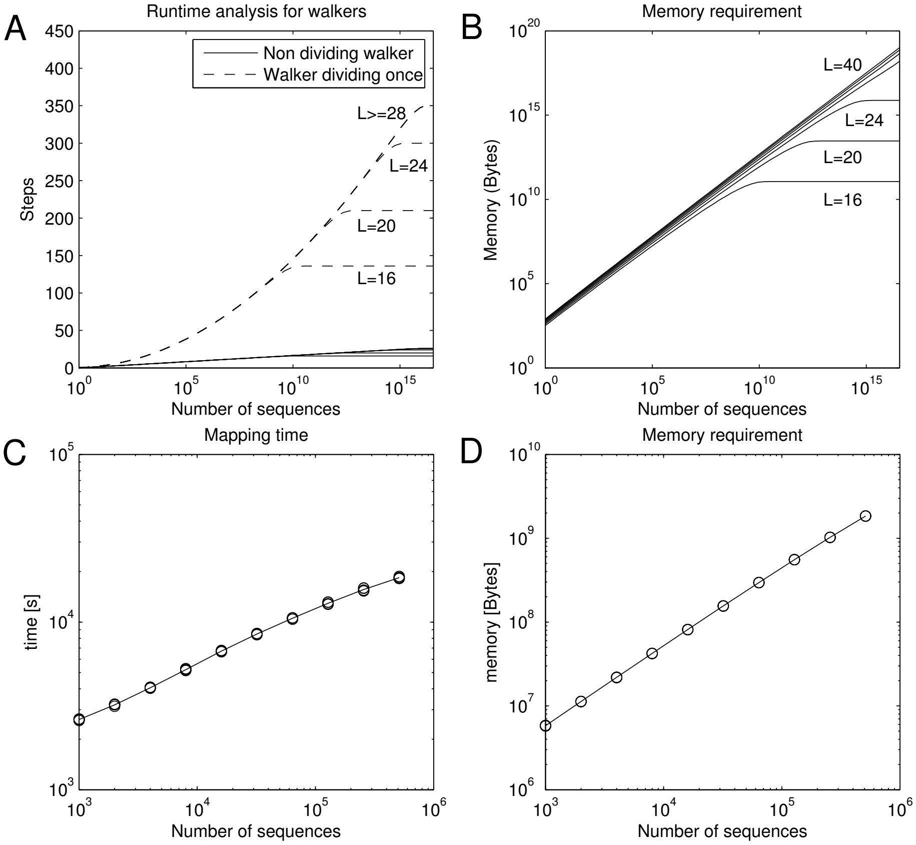


Figure 2.2: Performance of the mapping algorithm. A: estimated average number of steps performed by a walker as a function of the number of small RNAs represented in the tree. Solid line corresponds to the case of perfect matches only, dashed line to perfect and 1-error matches. The alphabet size was $A = 4$. The small RNA length varied from from 16 to 40, but the number of steps remains virtually unchanged for length > 28 nucleotides. B: estimated average memory requirements of the program as a function of the number of small RNAs in the input. The small RNA length varied from 16 to 40 nucleotides. C: Physical running time and D: memory requirements of the program on a 2.2 GHz AMD Opteron as a function of the number of small RNAs in the input. For each input size, we selected and mapped 5 random subsets of CAGE tags.

exhaustive list of all the alignment hits with up to one error. Since such an output can become huge in size, we recommend to use the -m command line argument to limit the number of hits per query sequence.

Command line arguments supported by oligomap:

1. -s scan only plus strand
2. -d scan all .fa target files in a directory
3. -r create a match report listing the number of hits for each query sequence
4. -m maximum hits to print for one query

Chapter 3

Inference of miRNA targets using evolutionary conservation and pathway analysis

This section was published in [49].

Dimos Gaidatzis, Erik van Nimwegen , Jean Hausser , Mihaela Zavolan*

Biozentrum,University of Basel,Basel,Switzerland
Swiss Institute of Bioinformatics, Basel, Switzerland
*Corresponding author

3.1 Abstract

MicroRNAs have emerged as important regulatory genes in a variety of cellular processes and, in recent years, hundreds of such genes have been discovered in animals. In contrast, functional annotations are available only for a very small fraction of these miRNAs, and even in these cases only partially.

We developed a general Bayesian method for the inference of miRNA target sites, in which, for each miRNA, we explicitly model the evolution of orthologous target sites in a set of related species. Using this method we predict target sites for all known

miRNAs in flies, worms, fish, and mammals. By comparing our predictions in fly with a reference set of experimentally tested miRNA-mRNA interactions we show that our general method performs at least as well as the most accurate methods available to date, including ones specifically tailored for target prediction in fly. An important novel feature of our model is that it explicitly infers the phylogenetic distribution of functional target sites, independently for each miRNA. This allows us to infer species-specific and clade-specific miRNA targeting. We also show that, in long human 3' UTRs, miRNA target sites occur preferentially near the start and near the end of the 3' UTR.

To characterize miRNA function beyond the predicted lists of targets we further present a method to infer significant associations between the sets of targets predicted for individual miRNAs and specific biochemical pathways, in particular those of the KEGG pathway database. We show that this approach retrieves several known functional miRNA-mRNA associations, and predicts novel functions for known miRNAs in nervous system development, inter-cellular communication and cell growth.

Our target prediction algorithm does not have any tunable parameters, and can be applied to sequences from any clade of species. It automatically infers the phylogenetic distribution of functional sites for each miRNA, and assigns a posterior probability to each putative target site. The results presented here indicate that our general method achieves very good performance in predicting miRNA target sites, providing at the same time insights into the evolution of target sites for individual miRNAs. The complete target site predictions as well as the miRNA/pathway associations are accessible on the ElMMo web server [50].

3.2 Background

Since the initial discovery of the lin-4 miRNA [51], and then of the let-7 miRNA which is highly conserved in evolution [12], combined experimental and computational approaches have resulted in the identification of hundreds of miRNAs in animal genomes, some of the large-scale studies being [4, 5, 29, 31, 39–41, 47, 52–60]. In contrast, high-throughput approaches for experimental identification of miRNA *targets* are only in their infancy [61, 62], and global properties of miRNA-dependent regulatory networks have mostly been inferred from computationally-predicted target sites [63–67].

Perhaps surprisingly, relatively little is known about the constraints on a functional miRNA target site. Mutational studies [68, 69] confirmed initial observations of Lai [70] and Lewis et al. [71] that perfect base pairing between the 5' end of the miRNA and its target is essential. As a consequence, some of the computational methods for miRNA target prediction require [63,72] or can enforce the constraint [73] that 6-8 nucleotides at the 5' end of the miRNA, the so-called miRNA "seed", are perfectly base paired with its mRNA target, or give a higher weight to the base pairs formed in this region [66, 74]. Since every 6mer occurs on average once every 4, 096 nucleotides in random sequence, the number of target sites for each miRNA would be very large if matching of the seed were the only requirement for functional target sites. Although there are indications that miRNAs do have a large number of targets [61–63, 69, 72, 75], experimental studies typically do not confirm that every seed match constitutes a functional target site. It seems therefore that additional factors contribute to the functionality of target sites. To improve the specificity of prediction of functional target sites, most computational studies make use of evolutionarily conservation [63, 66, 72, 76] or at least flag conserved putative targets [73, 74]. However, currently available methods generally use conservation statistics in an *ad hoc* manner. In particular, existing methods do not explicitly take the phylogenetic relationships into account when weighing the evidence of conservation between related species. In addition, current methods treat all miRNAs identically and ignore that the selection pressures for conserving functional target sites between related species may differ significantly between miRNAs. That is, functional target sites for one miRNA may be preferentially conserved in one subset of species, whereas the functional sites for another miRNA may be preferentially conserved in another subset of species. Incorporating conservation statistics in a general, rigorous and miRNA-dependent manner are the main features of the miRNA target prediction method that we present here.

From the very early stages of miRNA target prediction it became clear that regulatory proteins such as transcription factors are preferentially subjected to miRNA-dependent regulation. Yet, beyond a few well-characterized miRNA-target interactions, there is still very little known about the place of individual miRNAs in the regulatory networks of cells and organs. Several groups [64, 71, 77] have used Gene

Ontology categories in an attempt to characterize the biological roles of different miRNAs. Here we present a new analysis based on the association of targets for individual miRNAs with molecular pathways annotated in the KEGG database. This approach recovers some of the known miRNA-mRNA associations, and makes new predictions, in particular it predicts the for the involvement of specific miRNAs in nervous system development, inter-cellular communication, and cell growth.

3.3 Results and Discussion

3.3.1 miRNA-target interactions: the importance of different 'seed types'

Several lines of evidence [63,68–70,78] suggest that complementarity of the target site to the first 8 bases at the 5' end of the miRNA are of crucial importance for target site recognition. Lewis et al. [63] have investigated the importance of the miRNA "seed", defined as the positions 2-7 of the miRNA, by comparing conservation statistics of mRNA segments that are complementary to miRNA seeds with those of randomized control sets. They concluded that conserved 3' UTR regions predicted to hybridize perfectly with positions 2-8 of the miRNA or with positions 2-7 of the miRNA, but having an A nucleotide flanking the seed match at the 3' end are likely to be miRNA target sites. Inspired by these methods we decided to re-investigate the conservation statistics of different "seed types" across different clades of organisms.

In all cases, we only focused on the first 8 positions of the miRNA, and we analyzed the following 9 "seed types" (see Figure 3.1):

1. Perfect complementarity with Watson-Crick interactions between positions 1-8 of the miRNA and the mRNA target site.

2. Perfect complementarity at positions 1-7 but not at position 8.

3. Perfect complementarity at positions 2-8 but not at position 1.

4. Perfect complementarity at 2-7 but not at 1 and 8.

5. Complementary at positions 1-8 with a single G-U pair occurring with the U in the miRNA (GUM).

6. Complementary at positions 1-8 with a single G-U pair occurring with the U in the target (GUT).

7. Complementarity with a single bulged nucleotide on the miRNA side (BM).

8. Complementarity with a single bulged nucleotide on the target side (BT).

9. Complementarity with a single internal loop involving one nucleotide in both miRNA and target (LP).

For each of these 9 seed types t, and each of the four clades (mammals plus chicken, fishes, flies, and worms), we determined the fraction f_t of putative target sites that are perfectly conserved in all species of the clade (see Methods for details). We only considered miRNAs that were themselves conserved in all species of the clade. We also determined the "background" conservation fraction, of randomly chosen 3' UTR sequence segments of the same length as the respective seed types that are conserved in all species of the clade. Figure 3.1 shows the ratio of these two fractions, which we called "conservation fold enrichment".

As expected, octameric sites show most evidence of functionality. The conservation fold enrichment decreases dramatically as the extent of complementarity that we require between miRNA and putative target site decreases. In particular, sites in which only the nucleotides 2-7 of the miRNA are predicted to form base pairs with the mRNA, as well as sites predicted to form G-U base pairs or to contain internal loops show relative little evidence of conservation enrichment. This is not to say that such sites are never functional. Indeed, functional target sites of this type are known, in particular in worms [79], for which, interestingly, we observe the strongest evidence of selection on these seed types. However, for our target prediction method we decided to focus on the three seed types that show strong evidence of conservation enrichment across all clades: those with perfect Watson-Crick complementarity with positions 1-7, 2-8, or 1-8 of the miRNA. As a result, we predict the same set of target sites for different miRNAs with the same first 8 nucleotides. In reality, even though the seed is probably most important for targeting, the 3' ends may also contribute to the target

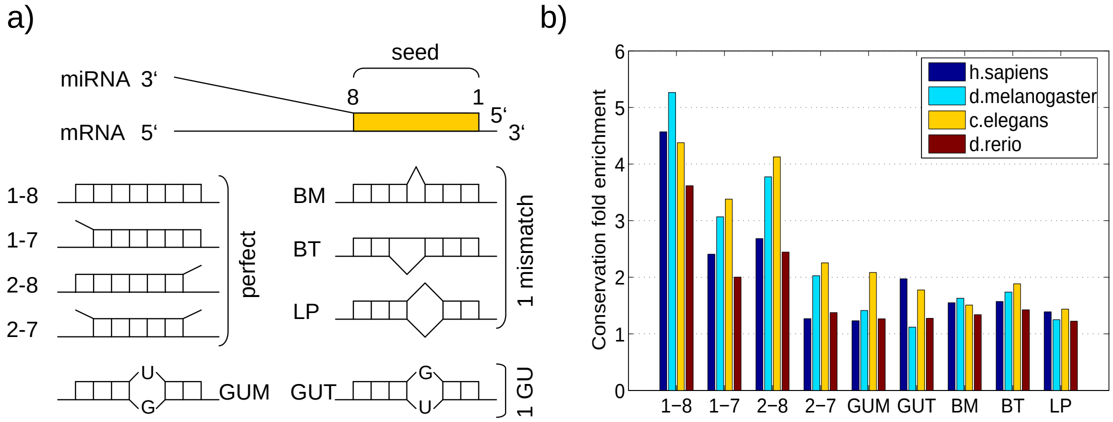


Figure 3.1: MiRNA seed types and conservation fold enrichment Schematic representation of the different "seed types" of miRNA target sites that we consider and conservation fold enrichment for each of them. a. Seed type interactions of miRNA-mRNA hybrids (see text). b. Conservation fold enrichment for the 9 different seed types in the four clades.

selection and this could differentiate the target sets for different miRNAs with the same seed. This possibility, which has been studied experimentally by Brennecke et al. [69], and is explicitly incorporated in other target prediction models [66], is not captured by our model. Note, however, that because the miRNA-mediated targeting depends on the expression of both miRNA and targets, distinct miRNAs that have the same seed sequence may still have different target sets simply due to differences in their expression profile, even though in principle they recognize the same set of target sites. Note also that we do not use a model in which the mRNA position corresponding to the first nucleotide in miRNA is an adenosine, because we did not find this constraint to consistently improve the conservation fold enrichment across all clades (not shown).

3.3.2 Bayesian phylogenetic model for miRNA target sites

We have developed a Bayesian probabilistic model for assigning, to each putative "site" in a 3' UTR that is complementary to a miRNA seed, a posterior probability that the site is a functional target site for the miRNA, meaning that the site has been selected in evolution for its ability to bind the miRNA. The details of the model are described in the Methods section, but the main ingredients are the following.

For each miRNA and each seed type t we collect all putative sites in the 3' UTRs of the reference species of the clade in question, i.e. the 3' UTR sequence segments that are complementary to the given miRNA seed. For each of these putative sites we then determine the conservation pattern $\vec{c}$, defined as a binary vector with $c_i = 1$ if the site is conserved in species i and $c_i = 0$ if it is not conserved in species i. We then count the number of times $n(\vec{c}, t)$ that conservation pattern $\vec{c}$ is observed for putative target sites of seed type t. To compute the posterior probabilities for individual sites, the model then compares these numbers $n(\vec{c}, t)$ with those that would be expected given the "background" frequencies $p(\vec{c}|t, \mathrm{bg})$ with which randomly chosen 3' UTR sequence segments of the same length as the miRNA seed show conservation pattern $\vec{c}$.

Generally, if conservation patterns with many $c_i = 1$ are much more abundant among putative miRNA sites than among background sites, then we infer that a fraction of the putative target sites must be functional and that selection has maintained these sites in some of the species. However, conserved target sites need not be functional in all species in which they occur. The conservation pattern of a given site is typically the result of selection maintaining the site in some of the species in combination with chance conservation of the site in other species, in particular those that are evolutionarily close. Our model flexibly and explicitly takes this into account. The model considers all possible "selection patterns" $\vec{s}$, which are also binary vectors, with $s_i = 1$ if the site is under selection in species i, and $s_i = 0$ if it is not. For each miRNA we then determine the frequencies $p(\vec{s})$ of different selection patterns that maximize the overall likelihood of the observed counts $n(\vec{c}, t)$. That is, we determine the distribution of selection patterns $p(\vec{s})$ that best explains the observed counts $n(\vec{c}, t)$ of conservation patterns for this miRNA.

Using the estimated frequencies $p(\vec{s})$ we can then determine, for each putative target site, the posterior probability that the site is functional given its conservation pattern $\vec{c}$. Finally to determine an overall probability that a given 3' UTR is targeted by a given miRNA we combine the posterior probabilities of all sites for the miRNA occurring in the 3' UTR. The reader is again referred to the Methods section for the details of all these procedures.

3.3.3 Phylogenetic distribution of functional target sites across miRNAs

Note that the estimated distribution over selection patterns $p(\vec{s})$ quantifies what fraction of putative sites in the reference species is under selection in each of the possible subsets of other species. That is, $p(\vec{s})$ estimates how functional target sites are distributed over the phylogenetic tree. Since we estimate $p(\vec{s})$ *independently* for each miRNA, our method allows us to compare how functional sites are distributed across the phylogenetic tree for different miRNAs. In Figure 3.2 we show the inferred phylogenetic distribution of functional target sites for 4 different human miRNAs. The genes of these miRNAs are conserved across all vertebrates shown in the figure. The precise parameters of the distributions $p(\vec{s})$ are represented by the bars at each node with red indicating the fraction of sites that remains under selection in the left descending branch only, blue the fraction that remains under selection in the right descending branch only, and green the fraction that remains under selection in both descending branches. The distribution $p(\vec{s})$ is also summarized in the thickness of the branches of each tree. Starting from the root (human) the thickness of each branch indicates what fraction of functional target sites is under selection along that branch. Note that the initial branch leading away from human has the same thickness for all four miRNAs, meaning that the fraction of human target sites that is under selection in at least one of the other species is roughly equal for these 4 miRNAs. However, as the figure shows, the two miRNAs on the left and the two miRNAs on the right differ significantly in the inferred pattern of selection across the tree. In particular, whereas the target sites for the miRNAs on the right (miR-9 and miR-124a) tend to be shared between all mammals, and to some extent with chicken and opossum, the target sites for the miRNAs on the right (miR-544 and miR-205) are shared mostly among primates, but not with other mammals. This suggests that, whereas the target repertoires of miR-9 and miR-124a have been largely conserved since the common ancestor of the mammals, significant changes have occurred in the target repertoires of miR-544 and miR-205 since that time. In Figure 3.3 we show the pa-

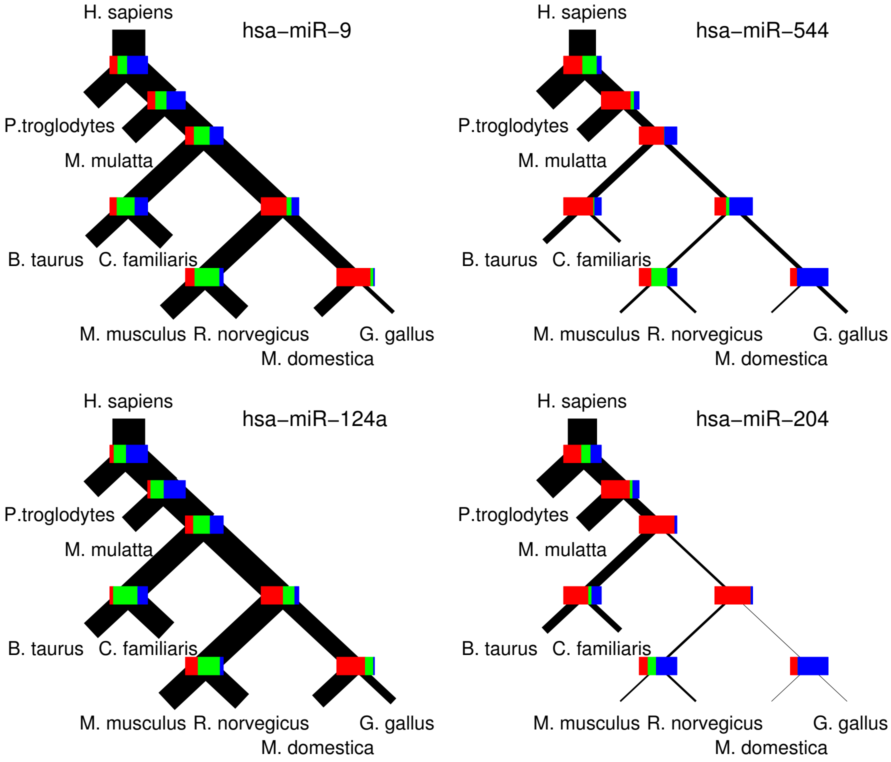


Figure 3.2: Examples of inferred phylogenetic distributions of functional target sites Comparison of the inferred phylogenetic distribution of functional target sites across vertebrate species (human - H. sapiens, chimp - P. troglodytes, rhesus maccaque - M. mulatta, mouse - M. musculus, rat - R. norvegicus, cow - B. taurus, dog - C. familiaris, opossum - M. domestica, chicken - G. gallus) for 4 different miRNAs. Starting from human at the root the thickness of the branches of the tree represents the fraction of putative target sites inferred to be selected along that branch of the tree. The bars at each internal node indicate what fraction of sites remains under selection in both descending branches (green), only the left descending branch (red), and only the right descending branch (blue). For each of the human miRNAs shown in this figure, there exists at least a miRNA with the same 1-8 “seed” sequence in all vertebrates in the tree.

rameters of the inferred selection distributions $p(\vec{s})$ for all miRNAs that are conserved across all warm-blooded vertebrates that we considered. These results provide a first comprehensive look into the species-specific targets of miRNAs.

Because we infer different distributions $p(\vec{s})$ for different miRNAs, and these distributions enter as priors in the Bayesian procedure, we generally assign different posterior probabilities to sites for different miRNAs, even if these sites have exactly the *same* conservation pattern. For example, in the example above a site for miRNA miR-544 that is only conserved in primates would get considerably higher posterior probability than a site for miR-9 with the same conservation pattern. This is because this conservation pattern corresponds better to the inferred selection pattern of miR-544 than the inferred selection pattern of miR-9.

One of the issues that has been extensively discussed in the miRNA literature is the question of the typical number of functional targets per miRNA, and the related question of what fraction of seed matches in 3' UTRs corresponds to functional target sites. Previous work has indicated that the number of targets per miRNA varies across miRNAs [64]. We believe that the ability of our method to infer species-specific miRNA targeting for each miRNA, allows for a more sensitive and accurate estimation of the total number of functional target sites of each miRNA.

There are two independent contributions to the total number of functional target sites for a given miRNA. First, the total number of miRNA seed matches varies from miRNA to miRNA and second, the fraction of seed matches that correspond to functional sites may vary from miRNA to miRNA. The latter can be estimated from the conservation evidence. In particular, the inferred parameter ρ of the distribution $p(\vec{s})$, corresponds to the fraction of miRNA seed matches that is under selection in the reference species *and* at least one of the other species in the clade. This provides a lower bound on the fraction of seed matches that is functional in the reference species (see Methods). By multiplying this fraction ρ by the total number of seed matches for the miRNA we obtain a lower bound on the absolute number of functional target sites for the miRNA. For simplicity we will refer to these as the estimated fraction of functional sites, and the estimated total number of functional sites. Figure 3.4 shows the estimated fraction of functional sites as a function of the estimated total number of functional target sites, for each clade of species and each miRNA. We infer that the

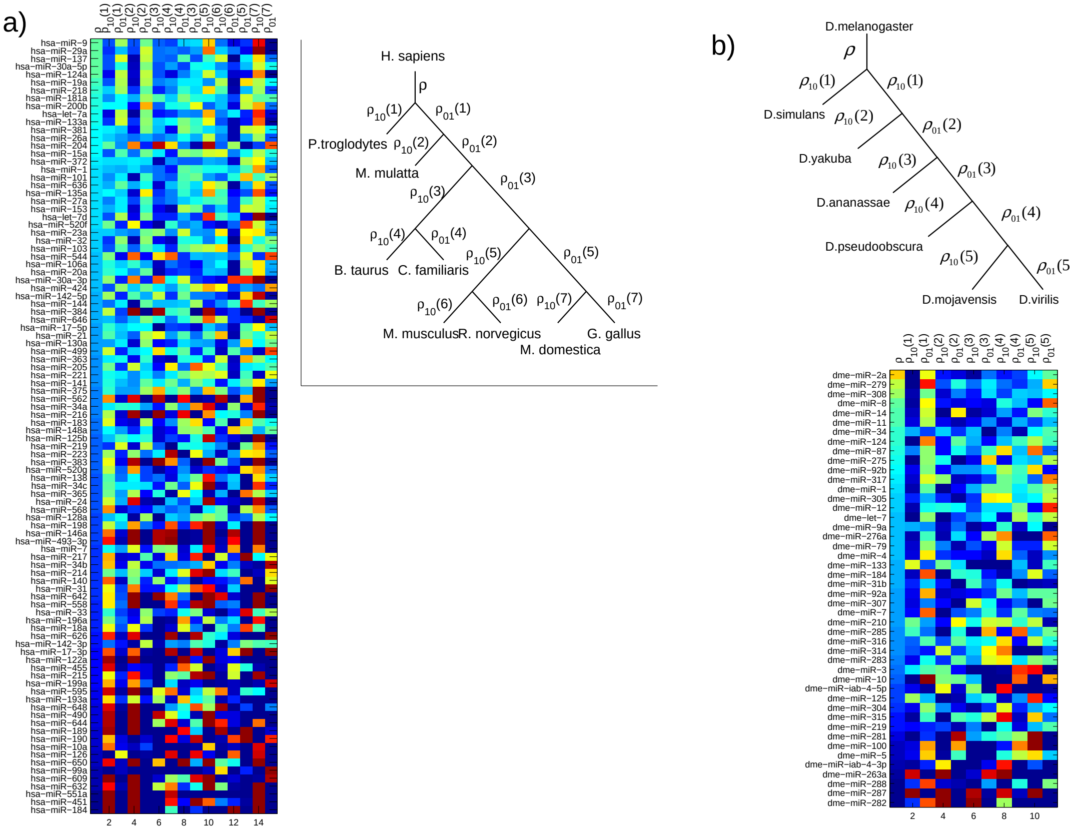


Figure 3.3: Phylogenetic distribution of functional target sites. Inferred selection pattern distributions $p(\vec{s})$ for all miRNAs that are conserved in all vertebrate (panel a) and all fly (panel b) species. Each row corresponds to a miRNA seed and each column corresponds to one of the variables $\rho_\omega(k)$ – where k indicates the internal node in the tree and ω indicates which of the subtrees are under selection – that parametrize $p(\vec{s})$ (see Methods). The miRNAs are sorted by the inferred total fraction ρ of putative target sites that is under selection in at least one other species.

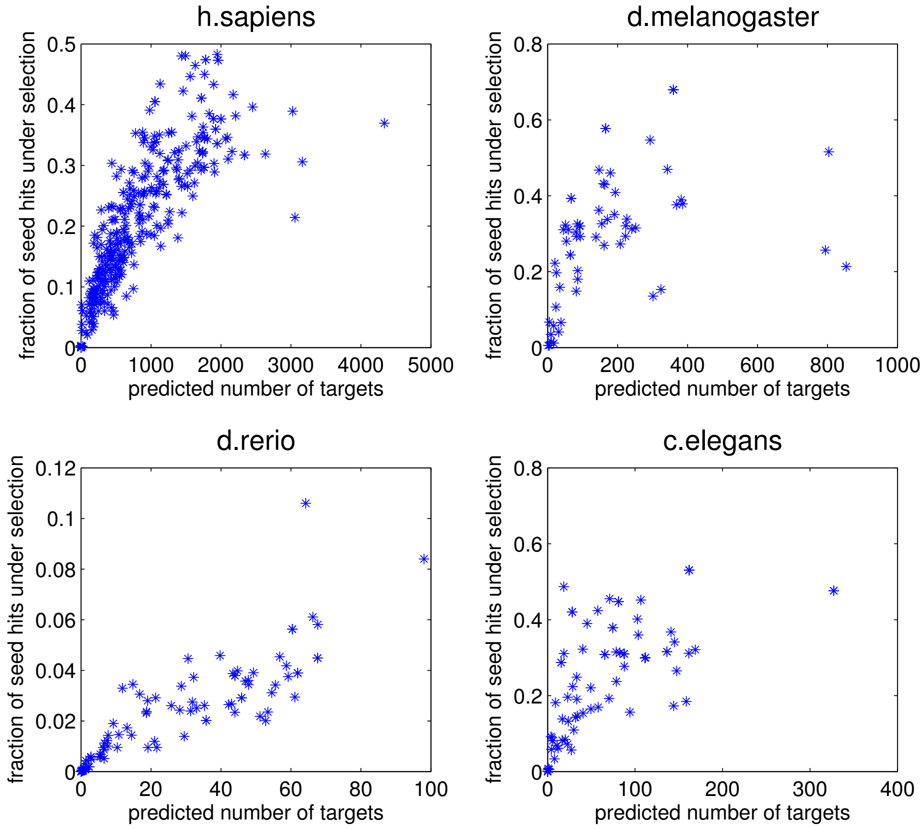


Figure 3.4: miRNA seed matches under functional selection The fraction ρ of seed matches inferred to be under selection (vertical axis) vs. the total number of sites inferred to be under selection in the entire set of mRNAs (horizontal axis) for individual miRNAs. Each star corresponds to one miRNA and each panel corresponds to one clade of species, with the reference species indicated at the top.

number of functional target sites varies very widely across miRNAs, i.e. from almost zero to several thousands. Similarly, the fraction of target sites under selection varies from close to zero to almost 50% in human, or even more in worms and flies. Overall we find that the average of ρ is about 30% for human, fly, and worm, meaning that we predict that at least 30% of miRNA seed matches in these species is functional. The inferred fractions ρ are significantly lower in fish. This is most likely because the reference species (Danio rerio) is relatively far (around 120 million years [80, 81]) from the other species in the clade (Fugu rubripes and Tetraodon nigroides), so that there is a smaller fraction of sites that is under selection in at least one of the other species. It is intriguing that, for all four clades, there seems to be a correlation between the fraction ρ and the inferred total number of functional sites at small values of ρ, but no correlation at high values of ρ.

The number of predicted sites under selection does not appear to correlate with the breadth of miRNA expression, as among the miRNAs with the largest number of predicted target sites we find some that are highly tissue specific (miR-9 and miR-124 that are expressed in the nervous system [39], and miR-155 that is specific to lymphoid cells [39]) as well some that have broad expression (e.g. the families of miR-29 [3, 53, 54, 82] and miR-30 [3, 82]). miR-16, which is ubiquitously expressed [3, 39] has an intermediate number of targets (Additional file 2 [49]).

3.3.4 Performance comparison with other methods

To assess the quality of our predictions relative to other methods that have been published to date, we built on the results recently published by Stark et al. [66], who have performed a detailed comparison of the performance of most of the prediction methods that are currently in use on a relatively large set of experimentally tested miRNA-mRNA interactions. This experimental data set has been mostly obtained by the Cohen lab, with a small number of interactions having been tested by other groups. The issues concerning the biases involved with the assembly of this data set have already been discussed by Stark et al. [66], and we will not belabor them here. We will only caution the reader that the accuracy of various different methods on this data set should not be taken as an indication of their accuracy on a random set of miRNA-mRNA interactions. Unfortunately, this unbiased experiment has not been done.

Since our method assigns a posterior probability to each predicted site, sets of predictions at different levels of confidence can be obtained by including only sites over a given posterior probability. We created such sets at different thresholds in posterior probability and computed the sensitivity ($\frac{TP}{TP+FN}$, i.e. the fraction of all true targets that were indeed predicted) and the specificity ($\frac{TN}{FP+TN}$) i.e. the fraction of all the correct negative predictions) for each set. The results are shown as the black line in Figure 3.5, which also shows the sensitivities and specificities of other prediction methods [73, 77, 77, 83–85], as inferred from the published results.

The figure indicates that our method performs as well as the most accurate prediction methods available to date, while maintaining a very high specificity even for high sensitivities. We observe a large overlap between our predicted targets and those predicted by Stark et al. [66] and Grün et al. [64] although there are also substantial numbers of predicted sites that are either specific to our method or specific to one of the other methods. The significant overlap is most likely a reflection of the similarity in the definition of target sites: 7/8-nucleotide seed matches that are conserved across at least some of the other flies account for a large fraction of the predicted sites in all three methods. However, these other methods also consider putative sites with fewer matches in the seed region if they are sufficiently conserved [64] or compensated by matches to the 3' end of the miRNA [66]. The very good accuracy of our predictions indicates that appropriately weighing the evolutionary information enables us to achieve a good performance even with more restrictive definition of putative target sites compared to thee other methods. In particular, we note that from a total of 12,155 high confidence predicted sites (posterior probability $p \geq 0.5$), a substantial proportion, namely 1,953 (16%), are not perfectly conserved in Drosophila pseudoobscura, but are conserved in many of the other flies. Such sites will be missed by methods that only consider strict conservation in D. pseudoobscura.

In Additional file 3 [49] we show a detailed comparison of our predicted target sets for fly miRNAs and those reported by Stark et al. [83] and Grün et al. [64]. We defined a UTR to be a predicted target of a specific miRNA if it had at least 0.5 probability of containing a functional site for the miRNA (see Methods). Because the UTR data sets used by different groups differs to some extent, we have used a conservative scheme of computing the overlap: we have assumed that whenever another method predicted a site in a splice variant of a given gene, all the variants would share the site. Thus, the numbers below represent upper bounds on the extent of overlap between the different methods. Note additionally that the total number of predictions made by other methods may not be the number of predictions reported in the respective studies, but include all the splice variants known to date. The overlap between our predictions and those of Stark et al. [83] and Grün et al. [64] varies significantly between miRNAs. For example, for the bantam miRNAs, which has shown to be involved in the regulation of cell growth [86,87], the overlap is quite large. We predict 140 targets of which 106 (76%) and 121 (86%) occur in the predictions of Stark et al. [66] and Grün et al. [64], respectively. The discrepancy is higher for miR-1, a miRNA required for muscle development [88]. We predict 362 targets of which 252 (70%) and 271 (75%) occur among the predictions of Stark et al. [66] and Grün et al. [64], respectively. Finally, for another microRNA, miR-281, we make only a total of 34 predictions of which only 13 (38%) and 17 (50%) occur among the predictions of Stark et al. [66] and Grün et al. [64], respectively. That is, at least half of our predicted targets are not predicted by the other two algorithms. Unfortunately, the data set of experimentally tested miRNA-mRNA interactions is too small to meaningfully compare the predictions of the different methods for individual miRNAs.

3.3.5 Location bias of predicted miRNA target sites in UTRs

We next turned to the high-confidence (posterior probability ≥ 0.5) subset of our predicted miRNA target sites and we asked whether we could identify a bias in the location of evolutionarily selected miRNA target sites in the 3' UTRs. Figure 3.6 shows a heat map representation of the location of these sites along the 3' UTRs in the different clades. In this plot, each predicted miRNA target site is represented as a dot with its x-coordinate being the total length of the 3'UTR in which it resides and its y-coordinate being the relative, normalized position of the site in the UTR. We infer that in all clades, the high-confidence sites tend to avoid the regions immediately after the stop codon as well as the end of the transcript. At the 3' end, this effect could be due to the presence of polyA tails in some of the Refseq transcripts. In human, where the UTRs are much longer than in the other species considered (3,300

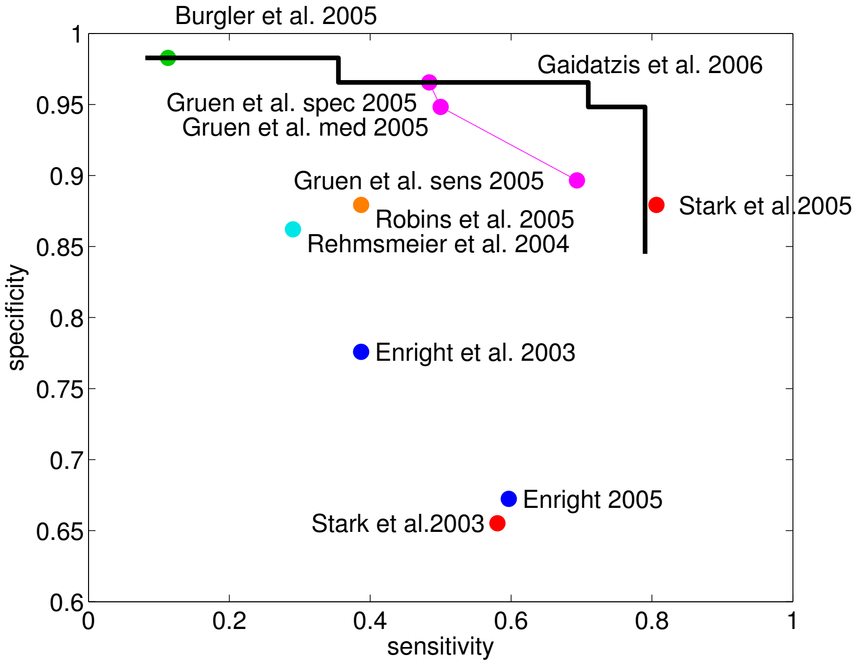


Figure 3.5: Performance comparison with other methods Comparison of the performance of our method and other published methods on a set of 120 experimentally tested miRNA-mRNA interactions in fly. Specificity (fraction of negatives that are not predicted) is shown as a function of sensitivity (fraction of positives that are predicted) for our method at different cutoffs in posterior probability (black line) and for other methods (colored dots).

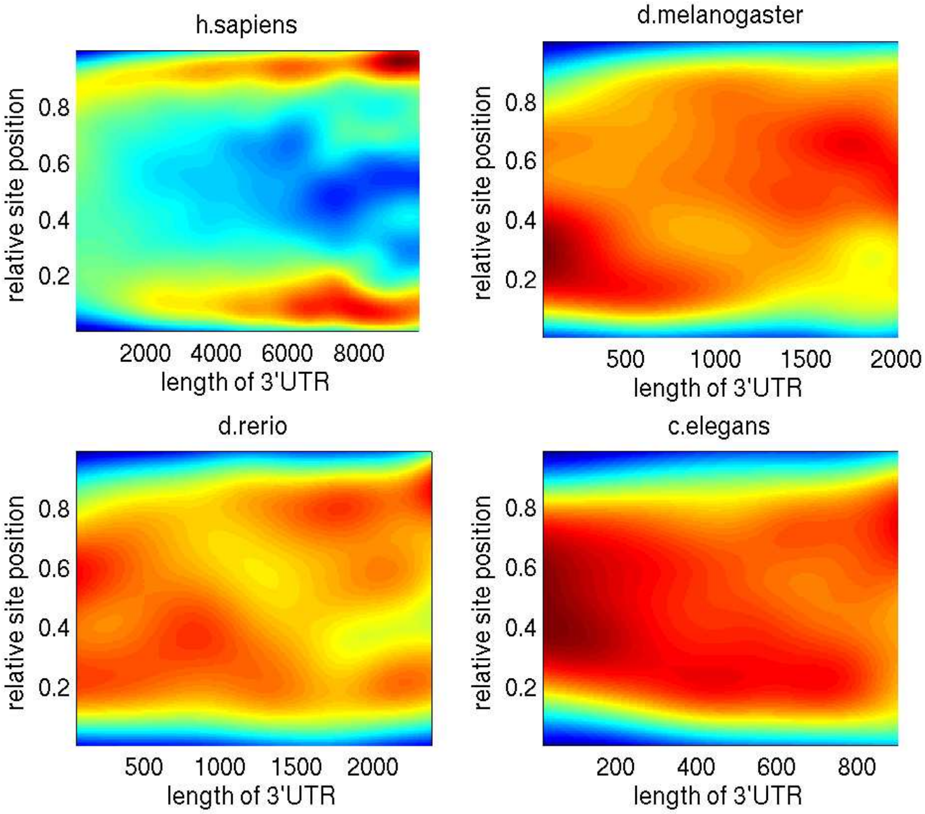


Figure 3.6: Location bias of predicted miRNA target sites in UTRs Distribution of predicted miRNA target sites in the 3'UTRs. Each predicted miRNA target site is represented by a dot with the x-coordinate corresponding to the length of the associated 3'UTR and the y-coordinate corresponding to the localization of the site within the 3'UTR normalized from 0 (start) to 1 (end). Gaussian kernels around all the dots were used to create a smooth interpolating density surface. Since the general UTR length distribution is not uniform, we normalized the vertical slices through the 2-D density surface $p(x, y)$ at each x-coordinate to obtain $p(y|x)$.

of the 22,459 of the human UTRs in our data set were longer than 2kb), conserved miRNA target sites also tend to avoid the regions in the middle of long UTRs. This pattern is mirrored in the conservation profile across long UTRs, i.e. long UTRs tend to be less conserved in the middle than toward their ends (data not shown). The pattern is also observed at the level of predicted target sites for individual miRNAs (Figure 3.7), i.e it is not caused by one or two miRNAs with an aberrant target site distribution.

A conceivable explanation for the observed pattern of enrichment of miRNA sites toward the start and end of long UTRs is that a non-negligible proportion of long Refseq UTRs erroneously contains introns. To test this hypothesis we obtained all EST sequences that overlap Refseq UTRs and calculated, for each UTR base, the fraction of all overlapping ESTs in which the base is intronic. As shown in Additional file 4 [49], there is almost no difference between the intron-inclusion profiles for long and short 3' UTRs. That is, the observed enrichment of miRNA sites toward the ends of long 3' UTRs cannot be explained by intron inclusion.

The observed pattern is interesting because it has been argued [66, 67] that miRNAs are a major factor driving the evolution of UTR lengths: ubiquitously-expressed genes have short UTRs, while genes whose expression is more restricted and regulated by miRNAs have longer UTRs. Our result suggests a more complicated scenario, in which more strongly conserved miRNA target sites, which have most likely emerged early, are located towards the boundaries of the 3' UTR, the stop codon and the polyadenylation site. This particular location of target sites may influence the likelihood of interaction between the miRNA-containing ribonucleoproteins and other complexes involved in RNA processing and regulation.

3.3.6 Inference of miRNA function using pathway analysis

To analyze the role that individual miRNAs play in the regulatory networks in human, we have used the KEGG database in which a large fraction of the human genes are assigned to pathways. KEGG provides a mapping between genes and pathways, as well as a reference to the identifier of each of the genes in the Gene database of NCBI. Based on this mapping, as well as on the assignment of Refseq identifiers to Gene identifiers which we obtained from NCBI, we have constructed an assignment

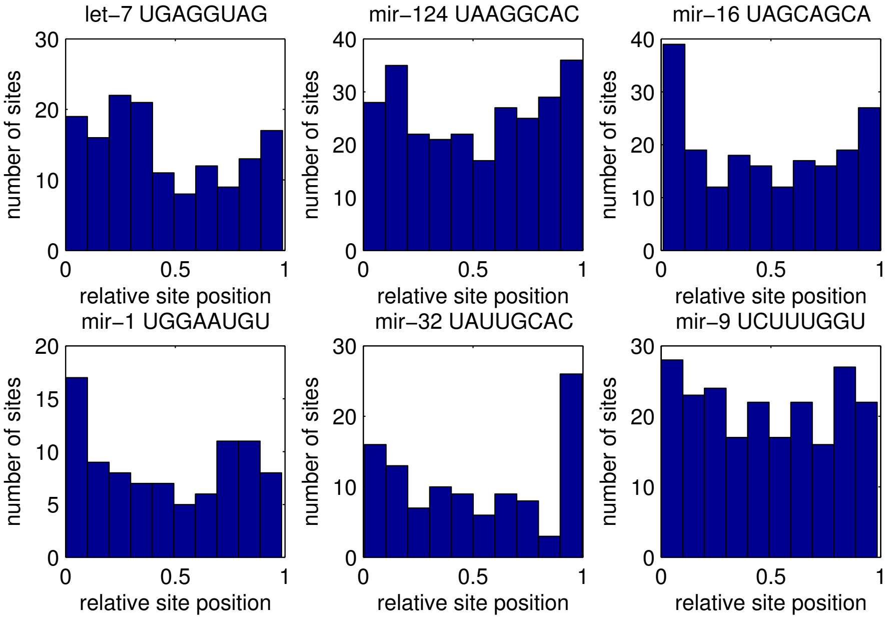


Figure 3.7: Location of predicted target sites of individual miRNAs in the 3' UTRs. Histogram of the relative position (0(start) to 1(end)) of high-probability predicted target sites (posterior probability $\geq$ 0.5) for 6 individual miRNAs in the human 3'UTRs longer than 4kb. The identity of the miRNAs and their corresponding seed sequences (positions 1-8 from the 5' end of the mature miRNA) are indicated on each panel.

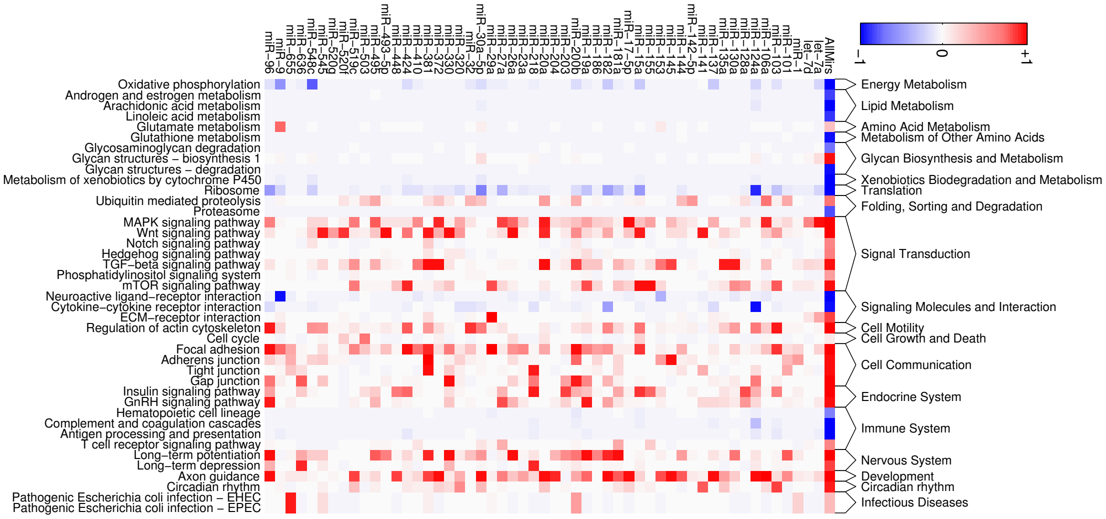


Figure 3.8: Pathway analysis. Representation of individual pathways among the predicted targets of a given miRNA. Each column corresponds to a KEGG pathway and each row to a group of miRNAs with the same seed sequence. Red indicates overrepresentation of the targets of a specific miRNA among the genes in the corresponding pathway, whereas blue indicates depletion. The intensity of the color indicates the posterior probability of the dependent model (see Methods). Pathways have been grouped in larger functional categories according to the KEGG annotation. Only miRNAs with at least one significant association are shown.

between putative miRNA targets and pathways. The resulting dataset consisted of 4, 011 human Refseq transcripts. Using putative target sites with posterior probability of ≥ 0.5, we have determined which pathways are significantly associated with each individual miRNA (see Methods). In particular, for each miRNA/pathway combination we calculated the log-likelihood ratio, given the observed data, of two models: one that assumes that pathway membership and being a predicted target of the miRNA are independent, and one that assumes that these are generally dependent properties.

Figure 3.8 shows the results of this analysis for the subset of miRNAs that had at least one significant association (Additional file 5 [49] shows the entire miRNA set). The color scale is centered around a log-likelihood ratio of 0 (white), and the intensity of the color is proportional to the posterior probability of the dependent model. Enrichment of targets in a pathway is shown in red, and depletion of targets in a pathway is shown in blue.

The first thing to note is that, as reported previously [66], genes that are ubiquitous and are involved in basic metabolic functions, tend to be depleted in miRNA target sites. Also noted before is that miRNAs tend to target genes involved in transcription regulation, intercellular communication, cell growth and death and development [63,64,66,71,74,77]. For example, we find that the targets of 19 of the 119 unique miRNA seeds are significantly enriched in the axon guidance pathway. This does not necessarily imply, however, that all these miRNAs are specifically involved in axon guidance. Many of the molecules involved in axon guidance are also involved in delivering spatial cues during the development of other systems, such as for example the cardiovascular system. So it is plausible that, whereas many miRNAs are associated with the axon guidance pathway, different miRNAs may act on different subsets of the mRNAs from this pathway in different tissues. Below we describe some of the most notable associations that we found between specific miRNAs and pathways.

Our method yields the expected associations for miRNAs which are specifically expressed in certain tissues (and presumably regulate processes that are specific to these tissues), or for miRNAs for which targets are already known. miR-124a, whose expression is highly specific to the nervous system, is one of the miRNAs most significantly associated with the axon guidance pathway. Its corresponding targets in this pathway include players with known involvement in nervous system development such as the ephrins B1, B2, and B3, ephrin receptors A2, A3, and B4, semaphorins 5A, 6A, 6C, and 6D, and plexins A3 and B2. As miR-124 is highly expressed in mature neurons, it is possible that its function is to maintain previously established neuronal circuits.

Our results also suggest an involvement of the miR-181 family of miRNAs in nervous system. These miRNAs, whose expression in zebrafish appears to be restricted to the nervous system, thymus and gills [56], have so far been shown to play a role in lymphocyte [89] and muscle [90] development. In our data, they have a set of high confidence targets in the long term potentiation pathway, among which glutamate receptors, calcium/calmodulin-dependent protein kinase II, adenylate cyclate 1, and calmodulin. In fact, calcium/calmodulin kinase II 2γ appears to play a role in both memory performance [91] and in activation-induced T cell differentiation [92]. These results may explain the up-to-now puzzling expression pattern of these miRNAs.

The let-7 miRNA, which was recently shown to regulate the let-60 gene in C.elegans and is presumed to regulate the human homologs of let-60, i.e. the Ras genes [76], is most significantly associated with the MAPK pathway, with the NRAS gene and the Ras guanyl releasing protein 1 RasGRP1 being predicted as high confidence targets. Additionally, let-7 is predicted to target several kinases and phosphatases in this pathway, and, importantly for the postulated involvement of let-7 in malignancy, the Fas ligand, TGFβ receptor I, nerve growth factor and fibroblast growth factor 11.

miR-9 has been described as a brain-specific miRNA [39], and recent evidence suggests that its expression is highest in fetal brain and oligodendrogliomas [93]. The top pathway associated with this miRNA is that of glutamate metabolism, in which miR-9 appears to target glutamate decarboxylase, glutamate dehydrogenase, glutamase, glutamate-cysteine ligase, glutamic-oxaloacetic transaminase 1, as well as glucosamine-phosphate N-acetyltransferase 1, 4-aminobutyrate aminotransferase, and phosphoribosyl pyrophosphate amidotransferase. The second most significant association for miR-9 is with with the focal adhesion pathway, in which many more genes appear to be targeted, among which collagen V α1, collagen IV α2, integrin 6, tenascin C, talin, trombospondin 2, and vinculin. These targets suggest that miR-9 may be involved in regulating the intercellular communication in the brain and the function of neural circuits.

Another group of miRNAs for which we suggest a role a development, in particular in the nervous system, is that of the embryonic miRNAs exemplified by miR-372, initially identified in a study of human embryonic stem cells [54]. These miRNAs appear to be primate-specific. However, the nucleotides at position 2-7, AAGUGC, are shared by the 5' ends of several other miRNAs that are embryonically-expressed and of restricted phylogenetic distribution such as the rodent miR-290 (AAAGUGCC 1-8), miR-291 (AAAGUGCU 1-8), miR-292 (AAAGUGCC 1-8), and miR-294 (AAAGUGCU 1-8), zebrafish miR-430's (U/AAAGUGCU at 1-8), as well as the human miR-302 group (UAAGUGCU at 1-8), miR-373 (GAAGUGCU at 1-8), and most miRNAs of the miR-520 group (AAAGUGCU at 1-8). Of these miRNAs, the study of [94] implicated miR-430 in the nervous system development in zebrafish, although in a subsequent study the authors showed that miR-430 plays a role in the clearance of maternal RNAs [2]. Our results speak to the first proposed role of this class of miRNAs, namely in nervous system development. The miR-372-related miRNAs (AAAGUGCU at 1-8) have a strong predicted association with the axon guidance pathway, where it is predicted to target, among others, the ephrin B2, ephrin receptors A4, A5 and A7, semaphorin 4B, LIM kinase 1, and p21-activated kinase 7. Moreover, at least some of the Smad genes that are part of the top pathway predicted to be targeted by these miRNAs, the TGFβ pathway, have been implicated in the growth of neurites [95]. Interestingly, the difference A vs. U or G at the first position between the miR-372 and other families mentioned above leads to quite different predictions of targeted pathways. For none of these other miRNAs have we found a pathway that appears to be significantly targeted.

Finally, we were very interested in understanding the function of miR-16 (which shares its seed with the miR-15 group of miRNAs), a miRNA that appears to be ubiquitously expressed at least in mouse [39], and has been implicated in regulation of apoptosis [96] and of mRNA stability [97]. We find that the most significant association of miR-16 is with the mTOR signaling pathway [98], which integrates nutrient-derived signals and controls cell growth. miR-16 appears to target the rapamycin-insensitive companion of mTOR, several ribosomal protein kinases, components of the eukaryotic translation initiation factor 4 (B and E), insulin-like growth factor 1 and others. The second most significant association of this miRNA is with the Wnt pathway, in which it targets several Wnt (Wnt2B, Wnt3A, Wnt5B, Wnt7A), a Wnt inhibitor (WIF1) and cyclin (D1, D2, D3) proteins, and the third most significant association is focal adhesion, where miR-16 appears to target a large number of transcripts that have fundamental role in cell division and cell-cell communication. Some examples are again the cyclins D1, D2, and D3, cell division cycle 42, p21-activated kinase 7 (PAK7), v-akt murine thymoma viral oncogene homolog 3 (AKT3), v-crk sarcoma virus CT10 oncogene homolog (avian)-like (CRKL), mitogen-activated protein kinase kinase 1 (MAP2K1), laminin gamma 1, B-cell CLL/lymphoma 2 (BCL2), and others. These suggest a fundamental role of miR-16 in controlling cell growth and maintaining cell-cell interactions. These functions may explain the observed association between miR-16/miR-15a deletions and chronic lymphocytic leukemia [99], and the slower progression of CLL in mice treated with rapamycin [100].

3.4 Conclusions

As the number of miRNA genes has been growing steadily, especially through high-throughput cloning techniques, the number of experimentally validated targets has been lagging markedly behind. Recently, studies that take advantage of the fact that miRNAs appear to also induce partial degradation of their mRNA targets have used microarray methodology to identify genes whose expression changes upon over-expression or knock-down of individual miRNAs. Typically hundreds of putative targets are identified in such studies but there is only partial overlap between these sets of putative targets and those that are computationally predicted using comparative genomics methods. Computational modeling of miRNA-mRNA interaction and accurate prediction of miRNA target sites therefore remains an important and challenging problem in bioinformatics. In particular, it is still poorly understood what constraints beyond matching of the miRNA seed determine functionality of putative target sites.

In this study, we developed a general method for miRNA target prediction that extends the already available methods in several ways. First, we treat the phylogenetic relationships between species in a rigorous and general way, without any tunable parameters. That is, the Bayesian procedure uniquely determines the posterior probabilities for each conservation pattern and seed type in terms of the observed conservation patterns of target sites for each miRNA. Thus, in contrast to many other target predictions methods which are specifically tailored to operate on a particular clade of species, our method can be applied to any clade of species, and the phylogenetic relations between the species will be automatically taken into account when assessing the significance of the site conservation patterns. This will, for example, enable us to easily update our predictions as more genomes become available, without the need of adapting the method.

Note also that our Bayesian procedure for incorporating information from conservation statistics is generally independent from the “site” definition that we employ and can easily be applied to other target site definitions (see Methods for details). Thus, if a better definition of target sites is developed in the future, for example through a better understanding of the requirements on functional miRNA target sites, then we can easily adapt the method to incorporate conservation statistics in essentially the same way. Most generally put, given a binary function that distinguishes “sites” from “non-sites” in RNA sequences, and given a set of “background frequencies” $p(\vec{c}|bg)$ with which sites defined by such a function show conservation pattern $\vec{c}$ by chance, we can apply the same methodology to assign posterior probabilities to all putative sites, incorporating the information from the conservation statistics of these sites.

Second, we estimate the evolution of selection pressures on target sites in a miRNA-specific manner. This enables us to correctly treat miRNAs that appeared at different stages in evolution, and whose targets may have undergone different selection pressures in different lineages. In particular, we show that different miRNAs show markedly different distributions of functional target sites across the phylogenetic tree and provide the first comprehensive picture of species-specific and clade-specific miRNA targeting. We have additionally shown that, especially in long 3’ UTRs that occur in vertebrates, miRNA target sites show a significant bias toward occurrence near the start and end of the 3’ UTR. This suggests the possibility that the choice of a distal polyadenylation site may reduce the activity of a miRNA target cassette in the center of the 3’ UTR, while introducing other miRNA target sites close to the new polyA tail.

With respect to the performance of our algorithm, we have shown that in fly, where extensive comparisons of the performance of target prediction algorithms have been done, our method performs at least as well as the most accurate methods available today, with a high specificity over a relatively large range of sensitivities.

Finally, to more robustly infer the function of individual miRNAs, each of whom may target hundreds of transcripts, we developed a method for identifying biochemical pathways that are significantly enriched or depleted in targets of a specific miRNA. We showed that, for well-studied miRNAs, this approach recovers the known functional associations. In addition, this analysis predicts novel pathway associations for a significant number of miRNAs.

3.5 Methods

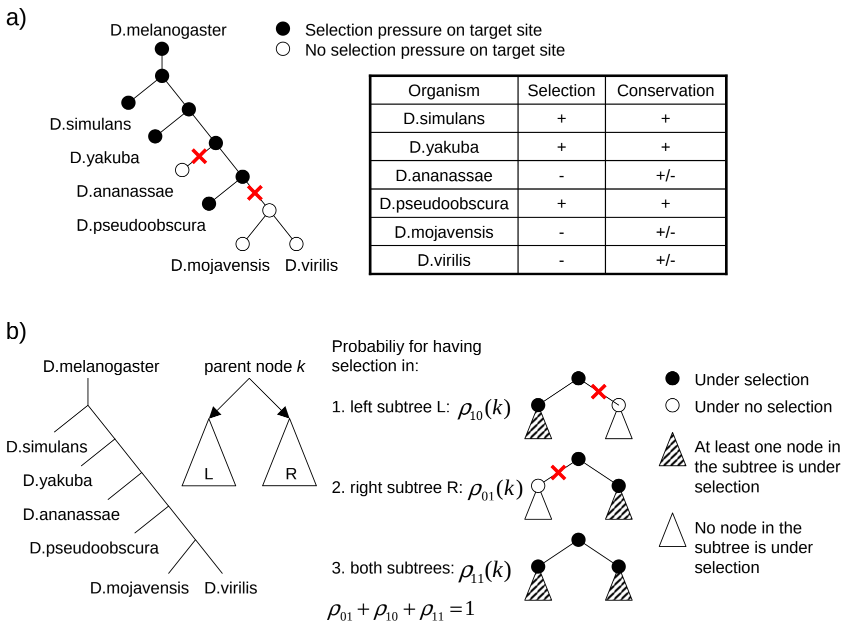
3.5.1 Conservation fold enrichment of different seed types

For the data shown in Figure 3.1 we focused, for each clade, on all miRNAs that occur in all species of the clade. Given that the seed sequence is so important for our inference, we used small RNA cloning data in human to determine the most abundant form of each mature miRNA (Pfeffer et al. [13, 14] and M.Zavolan & T.Tuschl, unpublished data), and we used this form in our prediction (Additional file 6 [49]). To determine which miRNAs are conserved in the clade we started with miRNA genes annotated in miRBase and searched the genomes of the other species for matches to the mature miRNA. Whenever the mature miRNA mapped with at most one mismatch we consider the mature miRNA conserved in that species. Since our inferences only uses the first 8 nucleotides of a miRNA, we then consider a miRNA seed to be conserved in a species if there exists at least one mature miRNA in that species with the corresponding seed.

For each seed type t we located all sites in the 3' UTRs of the reference species that are complementary to a seed of type t for any of the conserved miRNAs and then computed the fraction c_t of these sites that are conserved in all other species of the clade. We also determined the “background” conservation frequencies b_t for each seed type by scanning all 3' UTRs of the reference species and computing the fraction of all sequence segments of the same length as the seed that are conserved in all other species of the clade. Note that all seeds of the same length have the same background frequency b_t. This is because we found that this frequency is largely independent of the number of occurrences of a particular sequence segment in the reference species. Finally, the conservation fold enrichment f_t of seed type t is defined as the ratio of observed and background conservation rates: $f_t = c_t/b_t$.

3.5.2 Bayesian phylogenetic miRNA target identification algorithm

For each miRNA and each of the three seed types we identify putative target sites separately and assign a posterior probability to each target site as follows. First we


a)
D.melanogaster
● Selection pressure on target site
○ No selection pressure on target site
D.simulans
D.yakuba
D.ananassae
D.pseudoobscura
D.mojavensis
D.virilis
b)
D.melanogaster
parent node k
D.simulans
D.yakuba
D.ananassae
D.pseudoobscura
D.mojavensis
D.virilis
L
R
Probabiliy for having selection in:
1. left subtree L: $\rho_{10}(k)$
2. right subtree R: $\rho_{01}(k)$
3. both subtrees: $\rho_{11}(k)$
$\rho_{01} + \rho_{10} + \rho_{11} = 1$
● Under selection
○ Under no selection
At least one node in the subtree is under selection
No node in the subtree is under selection

Organism	Selection	Conservation
D.simulans	+	+
D.yakuba	+	+
D.ananassae	-	+/-
D.pseudoobscura	+	+
D.mojavensis	-	+/-
D.virilis	-	+/-

Figure 3.9: Modeling the selection pressure on miRNA target sites. a. The phylogenetic tree of the species in the clade (here flies) is rooted at the reference species (here melanogaster) and selection is modeled starting from the root and moving down the tree (see Methods for details). At each internal node k there are probabilities for selection to be maintained in one or both children of the node (see Methods for details). b. Relationship between selection and conservation patterns: Example of a selection pattern on a particular set of orthologous target sites in flies. Open circles indicate absence of selection pressure, closed circles indicate presence of selection pressure. Selection pressure is absent in Drosophila ananassae, mojavensis and virilis (D.ananassae, D.mojavensis, and D.virilis). The possible conservation patterns consistent with the selection pattern for this target site are listed in the table. The site needs to be conserved in all species in which selection pressure operates, namely Drosophila simulans, yakuba and pseudoobscura (D.simulans, D.yakuba, D.Pseudoobscura). In the species in which selection pressure does not operate, the site may or may not be conserved.

find all “sites” that are complementary to the seed in the 3’ UTRs of the reference species. Using pairwise alignments between the reference species and the other species we determine, for each putative site, which other species have the site conserved. An individual site was considered conserved if all the base pairs predicted to form between the miRNA and this site in the reference species could also be formed with the corresponding sites, extracted from the genome alignments, in the other species. This defines a “conservation pattern” for each site, which is a binary vector $\vec{c}$ with $c_i = 1$ if the site is conserved in species i and $c_i = 0$ if the site is not conserved. For example, for the triplet of worms C. elegans, C. briggsae, and C. remanei, using C. elegans as the reference species, the vector $\vec{c} = (1, 1)$ indicates a C. elegans site that is conserved in both other worms, the vector $\vec{c} = (1, 0)$ a site conserved only in C. briggsae, the vector $\vec{c} = (0, 1)$ a site conserved only in C. remanei, and the vector $\vec{c} = (0, 0)$ a site conserved in neither of the other two worms.

The fact that a putative target site is conserved does not necessarily imply that the site is functional. Especially for closely-related species short sequence segments, such as the 7-mers and 8-mers of miRNA seeds, can easily be conserved by chance. This evolutionary dependency between orthologous sites can be taken into account in a number of different ways. For example, in RNAhybrid [73] the p-values for orthologous target sites are combined by fitting an “effective” number of orthologous sequences to the observed p-value distribution for randomly generated miRNAs. Here we aim to incorporate the conservation statistics in a Bayesian framework that takes the phylogeny of the species explicitly into account and recognizes that a conserved site may be under selection in any of the subsets of species in which the site is conserved. That is, to infer how likely it is that a given putative site with conservation pattern $\vec{c}$ is functional, we want to calculate how likely it is to observe this conservation pattern $\vec{c}$ given that the site is functional and has been maintained by selection in one or more species, and how likely it is to observe $\vec{c}$ in the absence of selection for maintenance of the site.

To this end we first define a “background model” that gives the probabilities $p(\vec{c}|t, \text{bg})$ to observe conservation pattern $\vec{c}$ “by chance” for a seed of type t, i.e. a particular 7-mer or 8-mer. By “conservation by chance” we mean that there is no specific selection for maintaining the complementarity of the region in question to the 5’ end of the miRNA. We did not, however, use a background model that simply reflects the probabilities to observe different conservation patterns under neutral evolution. Any particular putative target site may overlap or be part of a site that is functional for some other reason, and may therefore be more conserved than would be expected under neutral evolution alone. Therefore, to estimate the background probabilities $p(\vec{c}|t, \mathrm{bg})$ we calculated the overall frequencies with which all conservation patterns $\vec{c}$ occur in the alignments, averaged over all 8-mers for the 1-8 seed type, and averaged over all 7-mers for the 1-7 and 2-8 seed types. In previous work others [63] have estimated background frequencies of conserved seed matches independently for seeds that have different absolute frequencies in the 3' UTRs of the reference species. We, in contrast, only require the *relative* frequencies of different conservation patterns, and we have observed that these are largely independent of the absolute frequency of the seed match. Note that for a clade consisting of the reference species and g other species, we are estimating the relative frequencies of 2^g possible conservation patterns for each seed type. Further subdividing these 2^g different conservation patterns by the absolute frequency of the seed match would reduce the amount of data available per seed too much for an accurate estimation of all the parameters.

We next calculated how likely it is to observe different conservation patterns $\vec{c}$ given that the putative target site is functional in at least one of the species. To this end we had to quantify the effect of selection on functional target sites. This is very difficult to do in complete generality. For example, one would generally expect that mutations that destroy functional target sites can have wildly varying effects on fitness with some sites being almost lethal when destroyed and others having only very mild deleterious effects. In addition these fitness effects will generally differ from species to the species, even for orthologous functional target sites. Of course target sites can also be spontaneously created through mutations in 3' UTRs, and in some cases these will act as functional target sites that can have either beneficial or deleterious effects. Thus, the rates at which orthologous target sites appear and disappear through evolution is a complex function of fluctuating selection pressures of which we know virtually nothing.

In order to be able to calculate meaningful probabilities for observing different conservation patterns $\vec{c}$ for functional sites we therefore make the following simplifying assumptions. First, we assume that given a set of conserved putative target sites, each of the conserved sites can be either “functional” or “nonfunctional”. In this context “functional” means that selection has acted to ensure that the target site remains conserved and “nonfunctional” means that the target site has evolved according to the background model. To take the worm example, if a functional C. elegans site is functional in both other worms as well, than the site will necessarily be conserved in both, i.e. we will have $\vec{c} = (1, 1)$. If the site is functional in C. briggsae only, then we might observe either $\vec{c} = (1, 0)$ or $\vec{c} = (1, 1)$, because the site is necessarily conserved in C. briggsae, and it may still remain unmutated by chance in C. remanei.

Thus, in general we consider all possible “selection patterns” for the site across the different species. Like the conservation pattern, a selection pattern $\vec{s}$ is a binary vector with $s_i = 1$ if the site is functional (under selection) in species i, and $s_i = 0$ otherwise. We calculate the probabilities $p(\vec{c}|t, \vec{s})$ to observe conservation pattern $\vec{c}$ given selection pattern $\vec{s}$ (and seed type t) as follows. Let $C(\vec{s})$ denote the set of all conservation patterns $\vec{c}$ that are consistent with the selection pattern $\vec{s}$. To be consistent with the selection pattern, the site needs to be conserved in all species in which it is presumed to be under selection, i.e. for all $\vec{c}$ in $C(\vec{s})$ we have that $c_i = 1$ for all i for which $s_i = 1$. The probability $p(\vec{c}|t, \vec{s})$ is then given by

$$p(\vec{c}|t, \vec{s}) = \frac{p(\vec{c}|t, \mathrm{bg})}{\sum_{\vec{c}' \in C(\vec{s})} p(\vec{c}'|t, \mathrm{bg})}. \tag{3.1}$$

Note that $p(\vec{c}|t, \vec{s})$ is just the probability that the site is conserved by chance in those species which have $c_i = 1$ but are not under selection, i.e. $s_i = 0$.

Finally, we need to quantify how likely it is *a priori* that a target site in the reference species will be under selection in a particular subset of the other species. That is, we need a prior probability distribution $p(\vec{s})$ that gives the probability that a miRNA site will be under selection in all species i for which $s_i = 1$. One of the key novel features of our model is that we allow this prior distribution $p(\vec{s})$ to vary between different miRNAs. We thus take into account the species- or clade-specific conservation of functional targets, i.e. that the reference species may share functional target sites with different subsets of species for different miRNAs.

For each miRNA we need to estimate the prior probabilities $p(\vec{s})$ of all possible selection patterns. That is, we need to estimate what fraction of putative sites in the reference species is under selection in each possible subset $\vec{s}$ of the other species. To do this we can first use the conservation of the miRNA *gene*. That is, if the miRNA *gene* is not conserved in a given species i, then we will assume that sites for this miRNA cannot possibly be under selection in species i. Thus, for every miRNA in the reference species we check which of the other species contains a miRNA with the same seed. When then set $p(\vec{s}) = 0$ for all vectors $\vec{s}$ in which the site is presumed under selection in a species that does not contain the miRNA. Note that, although unlikely, it is in principle conceivable that problems with the genome assembly of one of the species causes us to miss the ortholog of a particular miRNA gene. This will result in the conservation information from this species to be ignored for this particular miRNA.

The most general approach to estimating $p(\vec{s})$ would now be to simply find the distribution $p(\vec{s})$ that has overall maximum likelihood given the data. Formally, the probability $p(\vec{c}, t)$ to observe the conservation pattern $\vec{c}$ for a given putative target site of seed type t is given by summing over all possible selection patterns $\vec{s}$:

$$p(\vec{c}, t) = \sum_{\vec{s} \in S} p(\vec{c}|t, \vec{s}) p(\vec{s}), \tag{3.2}$$

where S is the set of all selection patterns that are consistent with the miRNA gene conservation pattern, $p(\vec{c}|t, \vec{s})$ is given by equation (3.1), and $p(\vec{s})$ is the prior probability distribution over selection patterns which we want to estimate. Let $n(\vec{c}, t)$ denote the number of occurrences of putative target sites of seed type t that have conservation pattern $\vec{c}$. The likelihood L given the data, i.e. the observed counts $n(\vec{c}, t)$, is then given by

$$L = \prod_{\vec{c}, t} p(\vec{c}, t)^{n(\vec{c}, t)}. \tag{3.3}$$

Given sufficient data, i.e. $n(\vec{c}, t) \gg 0$ for all $\vec{c}$, we could estimate $p(\vec{s})$ by maximizing L with respect to $p(\vec{s})$. The amount of data is limited, however, and the distribution $p(\vec{s})$ generally has a large number of independent components (2^g for g species). As we believe that it is not possible to robustly fit the entire distribution $p(\vec{s})$ without a significant risk of over-fitting, we instead aimed to parametrize reasonable distributions $p(\vec{s})$ using a much smaller set of parameters, i.e. on the order of g rather than 2^g parameters.

A second piece of information that can help us estimate $p(\vec{s})$ consists of the phylogenetic relationships between the species. That is, one would generally expect that functional target sites in the reference species are more often also functional in closely related species than they are in distantly related ones. It is thus natural to model the evolution of selection patterns along the branches of the phylogenetic tree of the clade. In analogy with evolutionary models for the evolution of gene sequences one might consider models in which selection for a site may "mutate" from "on" to "off" along each branch of the tree, with a probability of "mutation" proportional to the length of the branch. However, in contrast to such simple evolutionary events as point mutations in sequences, the "mutations" in our model correspond to changes in selection pressures and we see no reason to assume that these occur at a constant rate along each branch of the phylogenetic tree. Indeed, as we will see below, our results suggest that the rate of turnover of selection along a given branch of the tree differs significantly between miRNAs. To reasonable parametrize $p(\vec{s})$ we would therefore have to fit independent rates of loss and gain of selection along each branch of the tree for each miRNA. In addition, for every selection pattern $\vec{s}$ we would need to consider all evolutionary histories of selection loss and gain that are consistent with the resulting selection pattern at the leaves of the tree. Finally, note that we inherently treat the species in the clade asymmetrically. That is, we look for putative sites in the reference species only and then use pairwise genome alignments to determine the conservation pattern of each putative site in the reference species. We thus by definition never consider conservation patterns in which the site is conserved in some of the species but *not* in the reference species.

In summary, we looked for a parametrization of $p(\vec{s})$ that is flexible enough to allow for different rates of turnover of selection along each branch of the tree, that respects the topology of the phylogenetic tree, that takes into account our inherent asymmetric treatment of the reference species, and that minimizes the number of free parameters needed, so that over-fitting is avoided as much as possible. The parametrization that we chose is the following. We take the phylogenetic tree of the set of related species, and take the reference species as the root of the tree, as illustrated in Figure 3.9a for the Drosophila species. Starting from a functional site in the reference species we now move along the tree from top to bottom and assume that in each branch the "functionality" of the site can only be *lost.* That is, if the site is not under selection at a given internal node of the tree, we assume that it is also not under selection in any of its descendants. The probabilities $p(\vec{s})$ can then be parametrized by giving, at each node k, the probabilities $\rho_{11}(k)$, $\rho_{10}(k)$ and $\rho_{01}(k)$ that the functionality is maintained in both descendants, in the left descendant only, or the right descendant only (Figure 3.9b). Note that we assume that if the site was not under selection in either descendant then the site was already not under selection in the parent, and that at each node k the probabilities sum to one, $\rho_{11}(k)+\rho_{10}(k)+\rho_{01}(k) = 1$. There are thus 10 independent parameters for the Drosophila tree of Figure 3.9a which has 5 internal nodes. A final parameter ρ gives the probability that functionality is maintained in going from the reference species to the first internal node. Thus, with probability ρ the site is conserved in at least one of the other species, and with probability $(1 - \rho)$ it is specific to the reference species. The tree in Figure 3.9a shows a selection pattern with selection in D. simulans, D. yakuba and D. pseudoobscura. Using our parametrization the prior probability of this selection pattern is $\rho\rho_{11}(1)\rho_{11}(2)\rho_{01}(3)\rho_{10}(4)$ (we number the nodes from top to bottom). A nice feature of this parametrization of $p(\vec{s})$ is that the selection at all internal nodes of the tree is uniquely determined by the selection at the leaves of the tree, i.e. no sum over different evolutionary histories is required.

Note that by using only conservation information, we cannot possibly distinguish sites that are only functional in the reference species from sites that are not functional at all. That is, we do not know what part of the fraction $(1 - \rho)$ corresponds to sites that are functional, reference-specific sites and what fraction is nonfunctional. The inferred fraction ρ therefore provides a *lower bound* on the fraction of functional sites. For simplicity, we will make the conservative assumption that only the fraction ρ of sites is functional, and refer to these sites as the fraction of "functional" sites.

For each miRNA we estimate the parameters ρ, and $\rho_{11}(k)$, $\rho_{10}(k)$ and $\rho_{01}(k)$ for each node k, by maximizing the likelihood of the distribution given the observed data, i.e. equation (3.3). Let ω denote one of the possible selection patterns for the two descending branches, i.e. $\omega \in \{01, 10, 11\}$, and define the indicator function $\delta(\vec{s}, \omega, k)$ such that $\delta(\vec{s}, \omega, k) = 1$ whenever the parameter $\rho_\omega(k)$ occurs in $p(\vec{s})$ and $\delta(\vec{s}, \omega, k) = 0$ when it does not. We then have for the derivatives

$$\frac{dp(\vec{s})}{d\rho_\omega(k)} = \delta(\vec{s}, \omega, k)\frac{p(\vec{s})}{\rho_\omega(k)}. \tag{3.4}$$

Using this it is easy to show that L can be maximized with respect to the parameters $\rho_\omega(k)$ by an expectation maximization (EM) procedure. If we define

$$X_\omega(k) = \sum_{\vec{c},t} n(\vec{c},t) \left[\sum_{\vec{s}\in S} \delta(\vec{s},\omega,k) \frac{p(\vec{c}|t,\vec{s})p(\vec{s})}{\sum_{\vec{\sigma}\in S} p(\vec{c}|t,\vec{\sigma})p(\vec{\sigma})} \right] \tag{3.5}$$

then the EM update equations are given by

$$\rho_\omega(k) = \frac{X_\omega(k)}{\sum_{\tilde{\omega}\in\{01,10,11\}} X_{\tilde{\omega}(k)}}. \tag{3.6}$$

By iterating these equations we can determine the optimal $\rho_\omega(k)$. Since, as can also be shown by taking second derivatives, the likelihood L is a concave function of the parameters $\rho_\omega(k)$, the EM procedure is guaranteed to converge to the unique global optimum of the likelihood.

Once all $\rho_\omega(k)$ have been determined for a given miRNA we can calculate posterior probabilities of functionality for each putative target site as follows. As mentioned above, we consider a target site functional if it is under selection in the reference and at least one other species. The nonfunctional sites are then by definition those sites that are not under selection in any of the other species. We will denote this no-selection pattern as $\vec{0}$. For a site of seed type t and conservation pattern $\vec{c}$ the posterior probability that the site is functional is then given by

$$p(\vec{s} \neq \vec{0}|t,\vec{c}) = 1 - \frac{p(\vec{c}|t,\vec{0})p(\vec{0})}{\sum_{\vec{s}\in S} p(\vec{c}|t,\vec{s})p(\vec{s})}. \tag{3.7}$$

Note that the prior probability of no selection is simply $(1-\rho)$, i.e. $p(\vec{0}) = 1-\rho$, and that the probability for conservation pattern $\vec{c}$ given no selection is simply the background probability

$$p(\vec{c}|t,\vec{0}) = p(\vec{c}|t,\text{bg}). \tag{3.8}$$

We can thus also write the posterior as

$$p(\vec{s} \neq \vec{0}|t,\vec{c}) = 1 - \frac{p(\vec{c}|t,\text{bg})(1-\rho)}{\sum_{\vec{s}\in S} p(\vec{c}|t,\vec{s})p(\vec{s})}. \tag{3.9}$$

The parameter ρ thus corresponds to the estimated fraction of all putative target sites in the reference that are functional, i.e. under selection in at least on other species.

Finally, note that the sums in equations (3.1) and (3.2) involve a number of terms that grows exponentially with the number of species in the clade. We believe that this will not cause any computational problems in clades with less than 20 or so species. For much larger sets of species these sums can become computationally prohibitive. In those circumstances one could reduce the number of species by choosing, for each set of closely-related species, only a single representative. For species that are so closely-related that most putative target sites are conserved between them, choosing a single representative per group would hardly affect the predictions.

3.5.3 Sequence data

We carried out miRNA target predictions for all available human, fly, fish and worm RefSeq transcripts present in the 17th release of the Refseq database. We mapped all transcripts to the corresponding genomes using the Spa cDNA-to-genome alignment program [101], and the genome assemblies hg17 (human), dm2 (fly), ceWB05 (worm) and danRer3 (fish) provided by the Genome Bioinformatics group at the University of California, Santa Cruz [102]. From the same source we also downloaded pairwise alignments of several genomes with the genome of reference species, as follows: for human we downloaded hg17-to-panTro1, hg17-to-rheMac2, hg17-to-canFam2, hg17-to-bosTau2, hg17-to-mm7, hg17-to-rn3, hg17-to-monDom1 and hg17-to-galGal2; for fly we used dm2-to-droSim1, dm2-to-droYak1, dm2-to-droAna1, dm2-to-dp3, dm2-to-droMoj1, dm2-to-droVir1; for fish we used danRer3-to-fr1 and danRer3-to-tetNig1. Finally, for worm we used the software Threaded Blockset Aligner (TBA) [103] to align C. briggsae and C. remanei to C. elegans.

3.5.4 Pathway enrichment analysis

We used the KEGG database to infer pathways preferentially targeted by individual miRNAs. The KEGG database (ftp.genome.jp) contains mappings from NCBI Gene identifiers to pathway IDs (data files: [org]_ncbi-geneid.list, with [org] being the species code)), while the Gene database of NCBI (ftp://ftp.ncbi.nih.gov/gene/) provides mappings from Gene IDs to Refseq IDs (gene2refseq). By intersecting these data sets we obtained the mappings from Refseq IDs to pathways. We then used a Bayesian method to determine the significance of the overlap between the targets of each seed-equivalent set of miRNAs and each specific pathway.

For a given pathway and miRNA let n_{01}, n_{10}, n_{00} and n_{11} denote respectively the number of predicted targets of the miRNA that are not part of the pathway, the number of genes in the pathway that are not targeted by the miRNA, the number of genes that are neither targets of the miRNA nor members of the pathway, and the number of genes in the pathway that are predicted to be targeted by the miRNA. While pathway membership is a simple boolean variable (a gene is either a member of a given pathway or it is not), we can only assign probabilities for a given gene to be a miRNA target. Assume that a given gene has n putative target sites for a given miRNA and let p_i denote the posterior probability of the ith site. The probability that at least one of the sites is functional is then given by $p_{\text{tar}} = 1 - \prod_{i=1}^{n}(1 - p_i)$. We use p_{tar} as the probability that the gene is targeted by the miRNA and obtain n_{01} and n_{11} by summing p_{tar} over all genes that are not in the pathway and all genes in the pathway respectively. Similarly we sum $(1 - p_{\text{tar}})$ over all genes that are not in the pathway and all genes in the pathway to obtain n_{00} and n_{10} respectively.

Finally we calculate the probability of the observed counts n_{00}, n_{10}, n_{01}, and n_{11} under an "independent model", in which the probability to be targeted by the miRNA is independent of pathway membership, and a "dependent model" in which the probability of miRNA targeting is generally dependent on pathway membership. The likelihood under the independent model is given by

$$\begin{aligned} L_{\text{indep}} &= \int_0^1 (pq)^{n_{11}}(p(1-q))^{n_{10}} \\ &\quad ((1-p)q)^{n_{01}}((1-p)(1-q))^{n_{00}}dpdq \\ &= \frac{\Gamma(n_{1.}+1)\Gamma(n_{0.}+1)\Gamma(n_{.0}+1)\Gamma(n_{.1}+1)}{\Gamma(n+2)\Gamma(n+2)}, \end{aligned} \tag{3.10}$$

where $\Gamma(x)$ is the gamma function, a dot indicates summation over the variable in question, i.e. $n_{1.} = n_{10} + n_{11}$ and n is the total number of genes. For the dependent model the likelihood is given by

$$\begin{aligned} L_{\text{dep}} &= \int p_{00}^{n_{00}}p_{10}^{n_{10}}p_{01}^{n_{01}}p_{11}^{n_{11}}dp_{00}dp_{01}dp_{10}dp_{11} \\ &= \frac{\Gamma(4)\Gamma(n_{11}+1)\Gamma_{(}n_{10}+1)\Gamma(n_{01}+1)\Gamma_{(}n_{00}+1)}{\Gamma(n+4)}, \end{aligned} \tag{3.11}$$

where the integral is over the simplex $p_{00}+p_{10}+p_{01}+p_{11} = 1$. The ratio of likelihoods $L_{\text{dep}}/L_{\text{indep}}$ quantifies the amount of evidence for association between the miRNA targets and the pathway. This association can either be positive (miRNA targets are enriched in the pathway) or negative (miRNA targets are depleted in the pathway). In Figure 3.8 we plotted the quantity $\mathrm{sign}(n_{11}n_{..} - n_{1.}n_{.1})p_{\mathrm{dep}}$, where $p_{\mathrm{dep}} = \frac{L_{\mathrm{dep}}}{L_{\mathrm{indep}}+L_{\mathrm{dep}}}$ is the posterior probability of the dependent model (assuming a uniform prior).

3.6 Authors contributions

DG, EvN, MZ contributed to all the stages of this project. JH implemented the web server. All authors read and approved the final manuscript.

3.7 Acknowledgements

DG and JH were supported by SNF grants 205321-105945 and 3100A0-114001 to MZ, and by the Swiss Institute of Bioinformatics.

Chapter 4

Off-targets of siRNAs

4.1 SiRNA off-targets resemble miRNA targets

4.1.1 Introduction

SiRNA off-target studies based on microarray data like [104] revealed the presence of siRNA off-target signatures. These are defined as sets of transcripts that are not the targets for which the siRNA was designed, yet are reproducibly downregulated upon transfection of a given siRNA. Subsequent studies [105,106] showed that a large fraction of these transcripts harbor siRNA seed complementary sites in their 3' UTRs, suggesting that siRNAs systematically enter the miRNA pathway and bind with their 5' ends to 3' UTRs. This opens, in principle, the possibility to learn about additional determinants of miRNA target site functionality, beyond the seed complementarity, from siRNA off-target data, which is available in larger amounts. From this point of view, we set to analyze 17 siRNA transfection experiments collected from [106–108]. Figure 4.1 documents that siRNAs exert a seed-dependent regulation by depicting mRNA response curves for transcripts with no, one or more seed hits in their 3' UTRs.

4.1.2 Identifying miRNA targets using microarray technology

Several lines of evidence suggest that a substantial fraction of human genes is subject to post transcriptional regulation by miRNAs. Indeed, a large fraction of the strongly conserved regions of human 3' UTRs are complementary to the 5' end ("seed") of miRNAs [63, 72], indicating that miRNAs exert an important selection pressure on the 3' UTRs. On the other hand, microarray studies [61, 62] suggest that only a 6-8 nucleotide complementarity between the 5' end of the miRNA and the target site frequently suffices for miRNA-dependent downregulation of mRNAs, and that many more, less conserved, targets exist. While there is a clear correlation between the presence of seed-complementarity sites in the 3' UTR and miRNA-dependent downregulation of mRNA levels [61, 62], this correlation is far from perfect. Many transcripts with seed-complementary sites do not appear to respond miRNA overexpression [61] or antagomir-dependent downregulation [62]. Figure 4.1 shows the response of various classes of mRNAs in the antagomir-122 transfection experiment (mir-122 knockdown) in mouse liver [62]. Transcripts that harbor a mir-122 seed complementary site in their 3' UTR have a higher propensity to be upregulated than transcripts with no seed hit. Still the peak of the distribution of the mRNAs that carry one or even more seed hits is centered around zero which indicates that most of the transcripts (80-90%) that have one seed hit change only slightly or not at all. This suggests that additional miRNA binding determinants exist beyond the presence of seed complementarity.

4.1.3 Positional bias of miRNA targets in 3' UTRs

In Chapter 3 we discussed a method to computationally identify miRNA targets based on extensive cross species conservation information. There we found that predicted miRNA target sites in human are not uniformly distributed across the 3' UTRs. For long 3' UTRs (greater 2000 nt) they tend to accumulate at the beginning (after the stop codon) and at the end of the transcript. This effect was not observed in 3' UTRs with lengths less than 2000 nt. To validate this hypothesis with an independent data set, we decided to analyze microarray off-target data from 17 siRNA

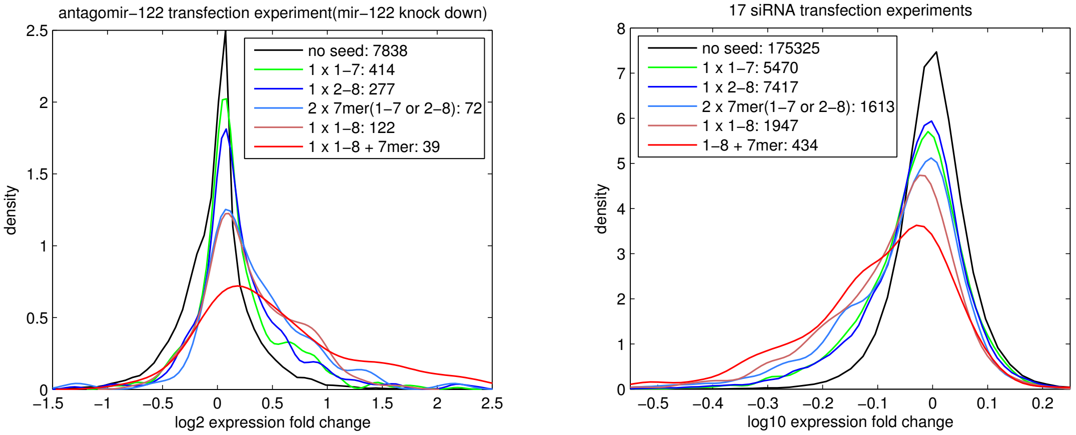


Figure 4.1: Distribution of the fold-change of various classes of mRNAs in the antagomir-122 transfection experiment(left panel) and 17 siRNA transfection expreriments [106–108] (right panel). The curves reflect the expression changes in subpopulations of transcripts which are defined by the presence or absence of certain types of seed complementary sites in their 3' UTRs. There exist 6 mutual exclusive subpopulations: transcripts with no seed hit, one 7mer seed hit to nucleotides 1-7 of the miRNA, one 7mer seed hit to nucleotides 2-8 of the miRNA, two 7mer seed hits (1-7 or 2-8), one 8mer (1-8) and one 8mer (1-8) plus one 7mer (1-7 or 2-8). The numbers close to the seed definitions reflect the number of transcripts within each group. Note that the right panel is a summary of 17 siRNA transfection experiments where all the datapoints were pooled together. Therefore the number of data points is much larger than in the left panel which just reflects one single experiment.

transfection experiments collected from [106–108]. We determined the off-targets of each individual siRNA by selecting all the transcripts that were downregulated with $p < 0.01$ on the microarray. Then we scanned the 3' UTRs of all the transcripts for words that were complementary to the seed region of the siRNA. We considered three types of seed binding, namely Watson-Crick base pairing to nucleotides 1-7, 2-8 and 1-8 of the 5' end of the siRNA. We collected siRNA off-target sites which were all the seed complementary sites located in 3' UTRs of transcripts that were downregulated ($p < 0.01$) upon siRNA transfection. To test for the location bias which we predicted to emerge at about 2000 nt we selected all the off-target sites that resided in 3' UTRs that were longer than 2000 nt and shorter than 4000 nt. We then monitored their position across the 3' UTR and compared them to the location bias that we computed from our predictions based on conservation (see Chapter 3). Figure 4.2 shows the results. Like predicted miRNA targets, siRNA off-targets determined experimentally seem to share this location bias in long 3' UTRs and tend to accumulate at the start and the end. Off-targets of siRNAs unlike miRNA target sites did not coevolve with the 3' UTRs so the fact that we observe this bias indicates that there could be a mechanistic reason for which RISC avoids the center part of long 3' UTRs such as e.g. mRNA looping between the stop codon and the end of the transcript.

4.1.4 Analysis of siRNA off-target site pairs

The analysis of miRNA/siRNA transfection data (Figure 4.1) showed that most of the transcripts that carry a seed match in their 3' UTR do not alter their expression (80-90%). Only about (10-20%) are significantly downregulated. We wondered whether the responding transcripts represent a random subset of the transcripts that carry seed-complementary sites, or whether the same subset of transcripts that carry seed complementary sites responds reproducibly in different experiments. To gain more insight we analyzed three siRNA transfection experiments from [106–108] for which we had two biological replicates, namely MAPK14-pos4-mismatch, MAPK14-pos5-mismatch and MPHOSPH1-202. The results are shown in Figure 4.3. Considering all transcripts, the Pearson correlation coefficient between the expression of mRNAs in the replicates varies between 0.49 and 0.77. When we only consider transcripts with at least one seed complementary site we detect correlations between 0.64 and

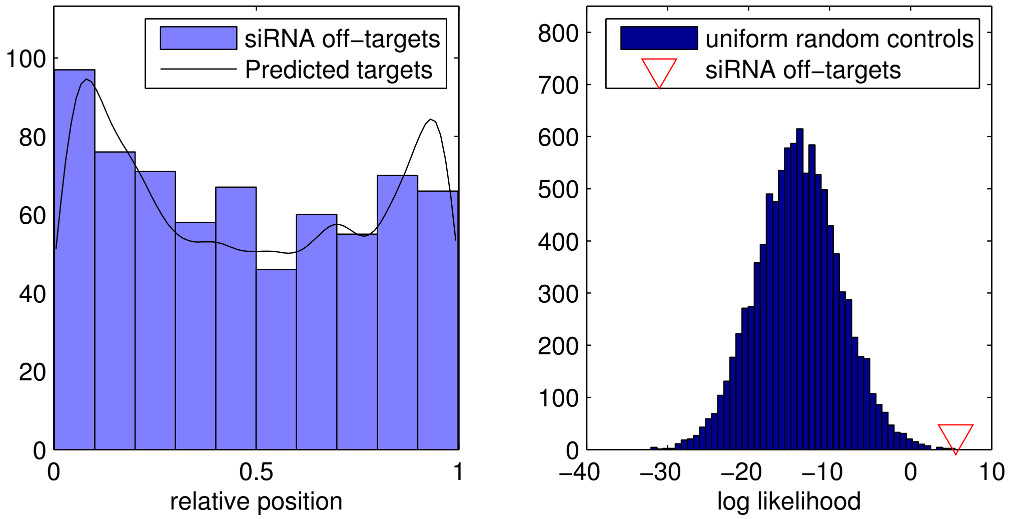


Figure 4.2: Positional bias of siRNA off-target sites. The left panels shows the density of sites along the 3' UTRs for predicted miRNA targets based on conservation $p_t(x)$ (black line) and siRNA off-targets determined by microarray experiments (histogram). To quantify the significance of the match, we generated random siRNA off-target sites x_i drawn from a uniform distribution and computed the log likelihood of the data given the distribution of sites computed from the predicted targets $l = \sum_i log(p_t(x_i))$. The right panel shows the log likelihood for 10000 randomly generated datasets. The red triangle denotes the score of the original siRNA off-targets.

0.86. Thus, the data suggest that the miRNA/siRNA target machinery does not pick its targets randomly from the pool of seed complementary sites, but rather that additional determinants contribute to reproducible targeting of the same transcripts.

Generally one could envision many different types of determinants. These might act globally on the whole mRNA level or locally for a given target site. Globally means that they control the overall potential of the mRNA to be targeted by the RISC. An example of such a determinant could be mRNA localization. Some mRNAs might be transported to a particular location in the cell where RISC has no access. As a consequence all the seed sites in their 3' UTR would not be functional and therefore would not respond in a transfection experiment. On the other hand, local determinants are defined to act in the immediate vicinity of a single target site. They could activate or inactivate single seed sites but their effect would not spread over the whole transcript. Examples for such determinants could be local secondary structure or auxiliary binding sites for instance for RNA-binding proteins. Whether the binding determinants are global or local can be determined because in the case of global effects all seed sites residing in the same 3' UTR should be either responding or not responding whereas in the case of local effects we expect seed sites located in the same 3' UTR to respond in a distance dependent fashion. One possible measure for this effect is the correlation of the fold changes for pairs of seed sites residing in the same 3' UTR. We plotted this quantity as a function of the distance between the two sites in a pair. The results are shown in Figure 4.4 (left panel). We can identify a clear peak for short distances which starts decreasing at a distance of about 150-200 nucleotides. This supports the idea that there might be local signals that act as enhancer elements. If two sites are located in the vicinity of such signals they will tend both to work. Two sites located in an unfavorable environment will tend both not to work. A close look at distances greater 150 nt reveals a somewhat puzzling plateau at the level of about 0.1 which is significant and persists over very long distances. This plateau argues for a global determinant because it means that no matter the distance between two sites they are somewhat coupled with respect to their response. Unfortunately we cannot determine what the nature of the global determinant is. It could be for instance the mRNA location, but it can also be that some mRNAs have been selected in evolution to be highly regulated by miRNAs, accumulating many

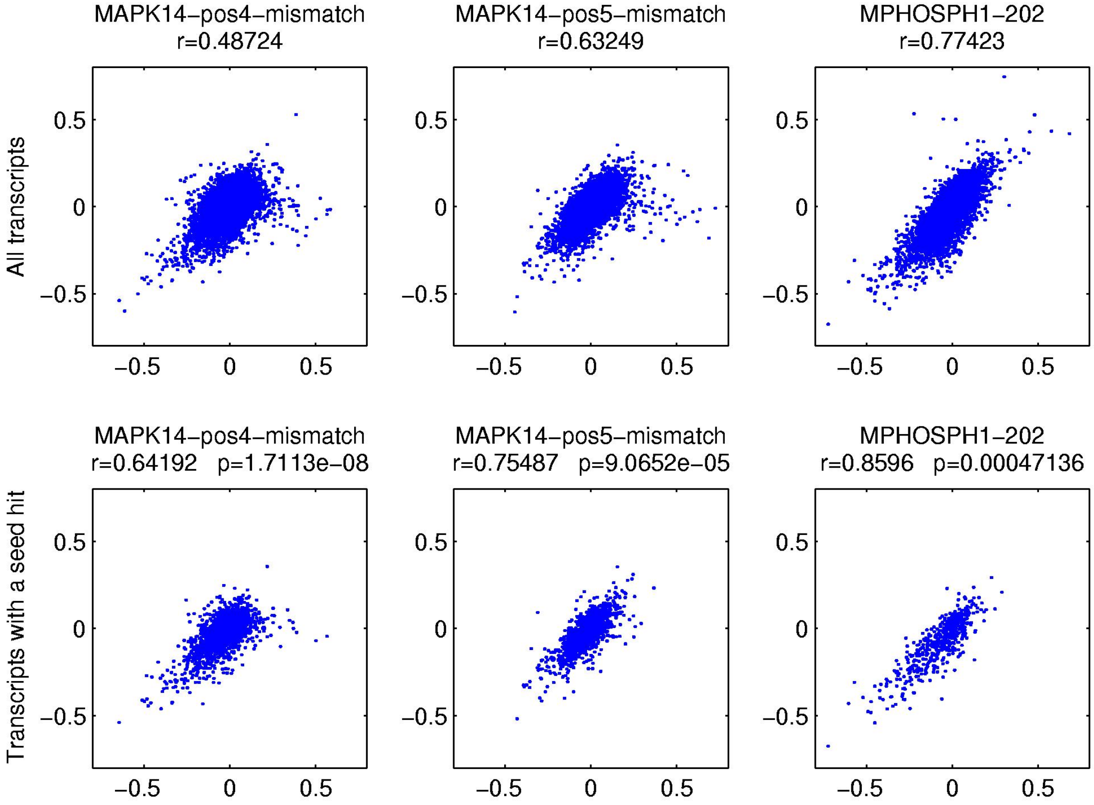


Figure 4.3: Reproducibility of siRNA transfection experiments. All panels depict scatter plots of expression fold changes for two independent siRNA transfection experiments. Panels 1-3 show all the transcripts that are on the array and that are expressed, panels 4-5 show only the subset of transcripts that contain at least one 1-7, 2-8 or 1-8 seed hit in their 3' UTR. The r values denote Pearson correlation coefficients. The p-values in the lower panels denote the significance of the increase in correlation with respect to the correlation for all the transcripts (upper panels).

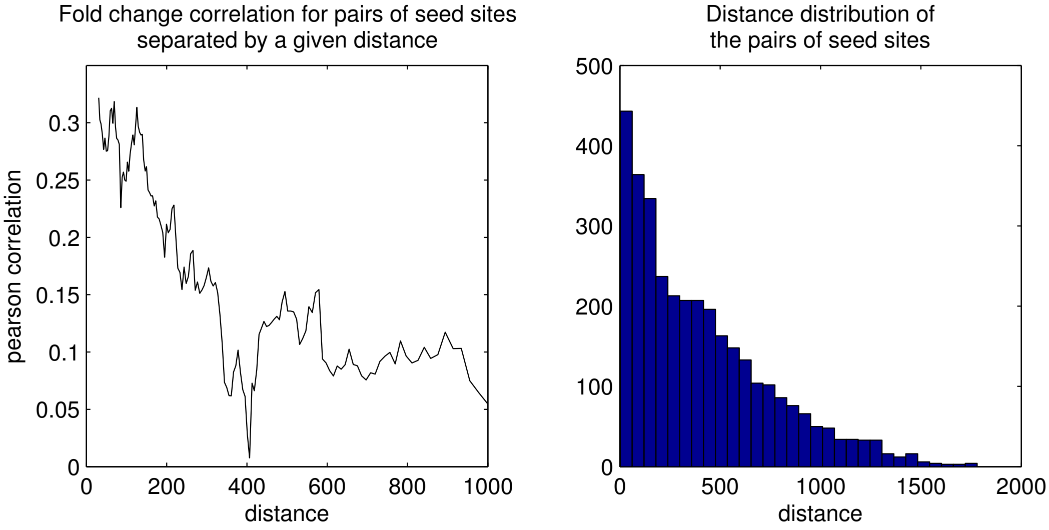


Figure 4.4: Correlation of the fold change for pairs of seed sites separated by a given distance. We determined all seed complementary sites to the 17 siRNAs for which we had microarray data [106–108]. We then enumerated all pairs of seed sites that were located in the same 3' UTR and recorded their relative distance. In the case where a 3' UTR contained two or more sites for the same siRNA we removed those sites from the pool. Furthermore we also removed all transcripts that were not expressed. Since those do not respond to the treatment they would induce global response correlations. Finally we also excluded transcripts that systematically respond in all of the 17 siRNA transfection experiments assuming that those changes are caused by the transfection procedure. We then sorted the pairs of sites (3375 in total) based on their distance. Using a window with the size of 500 we walked through this list of pairs and computed the Pearson correlation coefficient between the fold changes of the two pairs. The left panel shows these correlation as a function of the distance between pairs of sites (the distance is the average distance of all the pairs in a given window). The right panel shows a histogram of all distances present in the list of site pairs.

enhancer elements over their whole length. Identifying the molecular basis for these effects remains to be done in future work.

4.2 Strand-specific 5'-O-methylation of siRNA duplexes controls guide strand selection and targeting specificity

Parts of this section will be published in RNA 2008

4.2.1 Abstract

Po Yu Chen[1,2,#], Lasse Weinmann[2,#], Dimos Gaidatzis[3], Yi Pei[1,4], Mihaela Zavolan[3], Thomas Tuschl[1,*] & Gunter Meister[2,*]

[1]Howard Hughes Medical Institute, Laboratory of RNA Molecular Biology, The Rockefeller University, 1230 York Avenue, Box 186, New York, NY 10021, USA.
[2]Munich Center for Integrated Protein Sciences (CIPS), Max Planck Institute of Biochemistry, Am Klopferspitz 18, D-82152 Martinsried, Germany.
[3]Biozentrum der Universität Basel & Swiss Institute of Bioinformatics, Klingelbergstr 50-70, CH-4056 Basel, Switzerland.
[4]present address: RNA Therapeutics Merck & Co. Inc. , 770 Sumneytown Pike, WP26-410
[#]These authors contributed equally to this work.
[*]Corresponding authors

Small interfering RNAs (siRNAs) and microRNAs (miRNAs) guide catalytic sequencespecific cleavage of fully or nearly fully complementary target mRNAs or control translation and/or stability of many mRNAs that share 6 to 8 nucleotide (nt) of complementarity to the siRNA and miRNA 5' end. siRNA– and miRNA–containing ribonucleoprotein silencing complexes are assembled from double–stranded 21– to 23–nt RNase III processing intermediates that carry 5' phosphates and 2–nt overhangs with free 3' hydroxyl groups. Despite of the structural symmetry of a duplex siRNA, the nucleotide sequence asymmetry can generate a bias for preferred loading of one of the two duplex–forming strands into the RNA–induced silencing complex (RISC). Here we show that the 5' phosphorylation status of the siRNA strands also acts as an important determinant for strand selection. 5'-O-methylated siRNA duplexes refractory to 5' phosphorylation were examined for their biases in siRNA strand selection. Asymmetric, single methylation of siRNA duplexes reduced the occupancy of the silencing complex by the methylated strand with concomitant elimination of its offtargeting signature and enhanced off–targeting signature of the phosphorylated strand. Methylation of both siRNA strands reduced but did not completely abolish RNA silencing, without affecting strand selection relative to that of the unmodified siRNA. We conclude that asymmetric 5' modification of siRNA duplexes can be useful for controlling targeting specificity.

4.2.2 Introduction

Duplexes of 21–nt small interfering RNAs trigger RNA interference (RNAi) in mammalian cells and are widely used for functional genetic studies or screens in cultured cells (for reviews, see [109–113]). siRNA duplexes are designed to mimic the RNase III processing intermediates of naturally expressed dsRNAs, such as miRNAs, to effectively enter the RNAi pathway (for reviews, see [30,114–116]). Naturally processed siRNAs or miRNAs carry 5' phosphates and 3' hydroxyl groups and have symmetric 2–nt 3' overhangs ([117]). Synthetic siRNA duplexes with 5' hydroxyl ends are rapidly phosphorylated inside cells by the cellular kinase Clp1 ([118]). Some classes of small RNAs are additionally 2'-O-methylated at their 3' ends, depending on the species [119–126]. Mammalian miRNAs or siRNAs are not methylated, but the germline–specifically expressed piRNAs are 3'-end–modified [123, 124]. One strand of the siRNA duplex or miRNA/miRNA* molecule is assembled into an effector complex or RISC, while the other strand is degraded during the assembly process [127, 128]. The effector complex contains at its heart an Ago/Piwi protein member [128, 129]. Ago/PIWI proteins contain a conserved PAZ (Piwi–Argonaute–Zwille) and PIWI domain (for reviews, see [24, 130]). The PAZ domain, which is also present in Dicer, specifically binds the characteristic 2–nt 3' overhangs of RNase–

III–processed dsRNAs [131–134]. The PIWI domain contains a RNA 5'-phosphate binding (MID domain) and a RNase H domain [135–142]. The MID domain anchors the 5' end of the guide small RNAs [137–139], and presumably also plays a role during RISC–loading by receiving and binding the guide strand 5' phosphate [143]. Protein factors critically involved in siRNA or miRNA silencing complex assembly were first identified in Drosophila melanogaster. Duplex siRNAs are recognized by the heterodimer of RNase III Dcr-2 and the dsRNA–binding–domain protein R2D2, both of which are critical for formation of the Ago2–containing RISC [144, 145]. miRNA maturation in D. melanogaster is catalyzed by a heterodimeric complex of RNase III Dcr-1 and the dsRNA–binding–domain protein Loquacious/R3D1 [146–148]. R2D2 preferably binds the thermodynamically more stable end of the siRNA duplex and thereby directs strand selection [149]. The assembly of RISC is ATP–dependent, at least to a certain degree (for reviews, see [115, 150]). In mammalian systems, Dicer, the dsRNA–binding proteins TARBP2 and/or PACT, and an Ago protein appear to form the RISC–loading complex [151–155]. Two pathways are known for the transition of the duplex siRNAs or miRNA/miRNA* processing intermediate into a single–stranded–RNA–containing effector complex [156–158]. The first pathway requires near perfect base–pairing of the small RNA strands and depends on the RNase H activity intrinsic to a subset of the siRNA–binding Argonaute proteins [20, 135–140, 159–161]. RNase H active Ago proteins are able to receive the duplex siRNAs and guide the cleavage of the non–retained siRNA strand (often referred to as passenger, non–guide or sense siRNA) [156–158]. Upon release of the cleavage products, the retained guide (or antisense) siRNA is able to recognize complementary or partially complementary mRNA targets. The second RISC loading pathway is used, when duplex siRNA or miRNA/miRNA* duplexes either encounter a RNase–H deficient–Ago protein member, or when the duplexes are imperfectly paired across the center and cleavage site (like most miRNA/miRNA* duplexes) thereby preventing RNase H cleavage [156]. Presumably, a RNA helicase activity residing or transiently associating with the RISC–loading complex catalyzes the second RISC loading process [162–164]. The duplex–initiated RISC assembly process appears to be bypassed if high concentrations of single–stranded siRNAs are added to cell lysates or transfected into cells [165]. The role and the requirement for a 5' phosphate in reconstituting

RISC and its activity, however, remained somewhat controversial [136, 137, 159]. The specificity of small–RNA–guided mRNA degradation was examined in detail including mRNA array analysis [104–106, 166]. These studies revealed "off–targeting" activities of siRNAs that could not be separated from the "on–targeting" activity by simply decreasing the siRNA concentration. Some of the off–targets contained sequence segments of extensive complementarity to the siRNA, but many other off–targets showed only partial complementarity within their 3' untranslated region (UTR) to the siRNAs, notably at the 5' end of the siRNA guide strand. The latter observation was reminiscent of miRNA seed sequence (comprising pos. 1 to 8) mediated target mRNA regulation [61, 70, 71, 78, 83, 167]. Offtarget signatures can be identified for both the sense and the antisense siRNA strands, although strand biases during the assembly of RISC affect the targeting efficiencies of the two strands [168, 169]. In selecting siRNA sequences one takes now into consideration the differential thermodynamic stability of the siRNA ends to favor the incorporation of the target mRNA complementary guide siRNA (for review, see [170]. Other strategies proposed to control off–targeting activities included the introduction of 2'-O-methyl–ribose residues into the seed sequences of the siRNAs, which reduces off–targeting without detectable drop in on–targeting [107]. Here we study the role of the 5' terminal phosphate during RISC assembly from duplex and single–stranded siRNAs using 5'-O-methyl modified siRNAs. We show that the 5' phosphorylation status within a duplex siRNA is an important determinant of strand incorporation into RISC, and we demonstrate that selective 5'-O-methylation can be used to control strand–specific off–targeting activity. The phosphorylation status of single–stranded siRNAs has little impact on the non–natural RISC assembly and the subsequent activity of RISC.

4.2.3 5' phosphates are required for reconstitution of RISC from double–stranded but not single–stranded siRNAs

To revisit the requirements for 5' phosphates described for reconstituting RISC in D. melanogaster [143] or human cell lysates [136, 137, 159, 165], we prepared single– and double–stranded siRNAs with uridine and thymidine 5' end modifications (Figure 4.5). 5'-O-methyl–thymidine is currently the only nucleotide readily available for solid phase synthesis to render the ribose 5' ends of siRNAs refractory to phosphorylation in cell lysates [143]. HeLa cells and lysates contain hClp1 kinase, which rapidly phosphorylates 5'-hydroxyl termini of dsRNA or dsDNA as well as single–stranded RNA [118, 165]. To control for the concomitant introduction of a 5-methyl group with 2'-deoxy–thymidine incorporation into RNA, we also prepared siRNAs with 5'–hydroxyl–2'–deoxy–thymidine, 5'– hydroxyl–uridine, and 5' phosphorylated uridine–containing siRNAs.

HeLa S100 cell lysates were incubated with double–stranded siRNA derivatives followed by addition of 5' ^{32}P-labeled complementary target mRNA segments. Irrespective of the modification of the sense (passenger) strand, 5'-hydroxyl– or 5'-phosphate–modified antisense strands mediate target RNA cleavage. In contrast, 5'-O-methylated antisense siRNA showed substantially reduced activity (Figure 4.5B). The siRNA duplexes were cognate to firefly luciferase (Pp-luc) mRNA, and they were co–transfected with plasmids encoding the Pp-luc target and sea pansy control luciferase (Rr-luc) genes into HeLa cells. Consistent with the biochemical results, only the duplex with 5'-O-methyl–modified antisense strand showed reduced silencing activity (Figure 4.5C). Together, these observations were pointing to a role of the 5' phosphate of the antisense strand during RISC loading or RISC activity.

Two possibilities can be envisioned responsible for the reduced silencing activity of 5'-O-methylated antisense strand duplex siRNAs: (1) loading of the antisense strand into RISC was compromised and/or (2) the antisense–strand–loaded RISC had reduced activity because of conformational restraints imposed by an unoccupied Ago2 5' phosphate binding pocket [137–139, 159]. We therefore tested if we were able to load RISC using the single–stranded antisense siRNAs. Because single–stranded siRNAs are more susceptible to nucleases present in cell lysates than duplex siRNAs, we immunopurified FLAG/HA-affinity–tagged Ago2 protein complexes from HEK 293 cell lysates, and subsequently incubated them with single–stranded siRNAs and target RNA substrate. Surprisingly, the single–stranded siRNAs reconstituted RISC activity irrespective of their 5' modification status (Figure 4.5D).

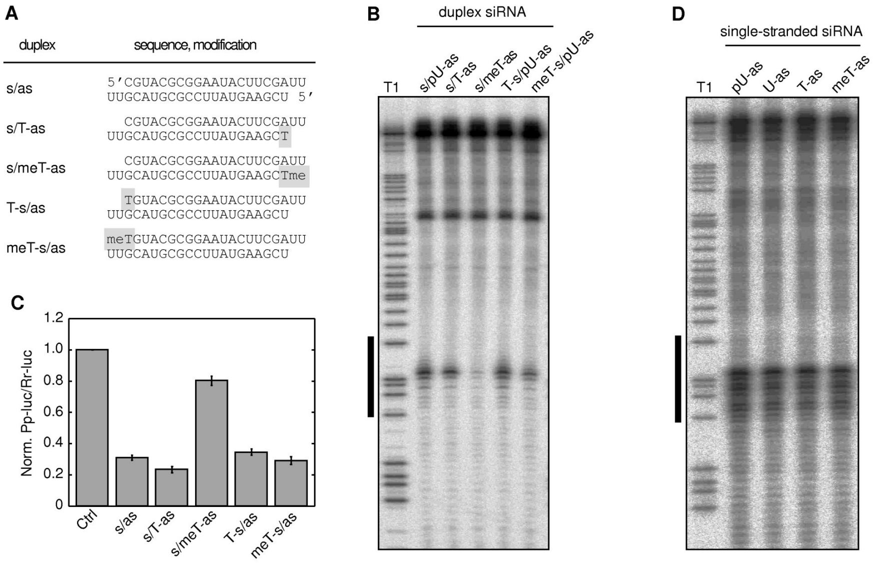
These data suggest that the 5' phosphate plays an important role during the process of RISC loading, that a 5' phosphate–sensing mechanism can be bypassed using single–stranded siRNAs and that the Ago2 5' phosphate binding pocket does not need to be occupied to mediate target mRNA cleavage.

4.2.4 Asymmetric 5'-O-methylation of duplex siRNAs directs strand selection during RISC formation

The loss of silencing activity of duplex siRNAs in which only the antisense strand was 5'-Omethyl- modified could be due to either preferential loading of the sense strand into RISC under these conditions, or to a defective recognition of the siRNA duplex by some RNAi machinery protein at a stage prior to RISC assembly. To monitor the asymmetry of siRNA strand incorporation and target RNA cleavage, we synthesized two pairs of siRNA duplexes that were predicted to be symmetrically and asymmetrically incorporated into RISC based on the differences in thermodynamic stability at their duplex termini [168, 169].

We first characterized biochemically the symmetric siRNA duplex (Figure 4.6A) by incubating it in lysates from HEK 293 cells transiently transfected with FLAG/HA-affinitytagged Ago2. HEK 293 cells were chosen because they are efficiently transfected at large scale with FLAG/HA-Ago2 expression plasmids. The siRNAs that co–immunoprecipitated with FLAG/HA-Ago2 were analyzed by Northern blotting using probes complementary to either the antisense or the sense strand. Signals for the antisense strand were detected when the siRNA duplex contained unmodified or 5'-O-methyl sense strand, but not when the antisense strand was 5'-O-methylated (Figure 4.6B). Signals for the sense strand were detected when the siRNA duplex contained unmodified or 5'-O-methyl antisense strand, but not when the sense strand was 5'-O-methylated (Figure 4.6B). We then confirmed that symmetrically or asymmetrically loaded FLAG/HA-Ago2 immunoprecipitates cleaved siRNA–complementary ^{32}P-cap–labeled target RNAs [171] as expected from their strand–loading ratios determined by Northern blotting (Figure 4.6C).

To measure the cell–based silencing activities from the assembly of the antisense and the sense siRNA strand into RISC, we introduced antisense– and sense–complementary sequence segments into the 3' UTR of EGFP as well as Pp-luc reporters. The results were consistent with our biochemical observations in HEK 293 lysates (Figure 4.7D, E). Symmetric 5'-O-methylation of both siRNA strands lead



duplex	sequence, modification
s/as	5'CGUACGCGGAAUACUUCGAUU UUGCAUGCGCCUUAUGAAGCU 5'
s/T-as	CGUACGCGGAAUACUUCGAUU UUGCAUGCGCCUUAUGAAGCT
s/meT-as	CGUACGCGGAAUACUUCGAUU UUGCAUGCGCCUUAUGAAGCTme
T-s/as	TGUACGCGGAAUACUUCGAUU UUGCAUGCGCCUUAUGAAGCU
meT-s/as	meTGUACGCGGAAUACUUCGAUU UUGCAUGCGCCUUAUGAAGCU

Figure 4.5: 5' phosphates are required for reconstitution of RISC from double–stranded but not single–stranded siRNAs. (A) Luciferase duplex siRNAs used in (B) and (C). (B) HeLa S100 extract was incubated with the siRNAs shown in (A). After preincubation, a ^{32}P-cap labeled RNA substrate was added and the cleaved RNA fragments were analyzed on a denaturing sequencing gel. T1 refers to partial nuclease T1 digestion. The black line to the left indicates the target site. (C) The effect of duplexes 1-5 and a control siRNA duplex on inhibition of the firefly luciferase (Pp-luc) expression relative to Renilla luciferase (Rr-luc) in a dual–luciferase assay. The ratios of the signals of Pp-luc/Rr-luc for duplexes 1-5 were normalized to that of the control siRNA(three independent experiments s.d.) (D) FLAG/HA-tagged Ago2 was transiently transfected into HEK 293 cells. Tagged proteins were immunoprecipitated from the lysates using anti–FLAG beads and RISC activity was reconstituted by adding singlestranded siRNA against the luciferase mRNA either with a 5' phosphate (lane 1), without a 5' phosphate (lane 2), with a 5' hydroxyl-2'-dT (lane 3) or a 5' methoxy-2'-dT (lane 4). p, 5' phosphate; Me, 5'-O-methyl group.

to an overall reduced gene silencing activity in the luciferase reporter assay, without changing the symmetry of the residual cleavage activity when compared with the unmodified siRNA duplex. In contrast, a single asymmetric modification did not alter the activity attributable to the unmodified siRNA strand. Together, these observations indicate that the RISC assembly of symmetrical siRNA duplexes can be influenced using asymmetric 5'-O-methylation, whereby the methylation of one strand directs incorporation of the complementary strand into RISC.

We next evaluated whether strand selection of a thermodynamically asymmetrical siRNA duplex could be controlled by 5'-O-methylation (Figure 4.8A). The activities of the modified and unmodified siRNA duplexes were determined using the EGFP and luciferase reporter assay described above (Figure 4.8B, C). The unmodified siRNA duplex preferentially repressed the target complementary to the sense strand, as expected from the design of the siRNA. The 5'-O-methylation of the antisense siRNA reduced its cleavage activity about 2- fold. Reciprocally, 5'-O-methylation of the sense siRNA strand reduced its cleavage activity 2-fold, while the antisense strand cleavage activity was significantly increased. Similar to the observations made for the symmetric siRNA duplex above, double 5'-O-methylation weakened the silencing activity of both the antisense and sense strands. These observations indicate that 5'-O-methylation of siRNA strands can influence siRNA strand incorporation into RISC, even in the context of a thermodynamically asymmetric siRNA.

4.2.5 Strand–specific 5'-O-methylation also controls siRNA off–targeting activity

Off–target effects are, like on–target effects, strictly sequence–specific and caused by either near–perfect complementarity between the central region of the siRNA and its targets or by seed–sequence complementarity between the siRNA and the target 3' UTR [104–106]. To assess if strand selection from asymmetrically 5'-O-methylated siRNA duplexes could be used for controlling siRNA offtargeting activity, we determined the gene expression profiles of HeLa cells transfected with 5'-O-methylated or unmodified siRNA duplexes. The symmetrical siRNA duplex, whose both sense and antisense strands are incorporated into RISC was selected for the analysis (Figure

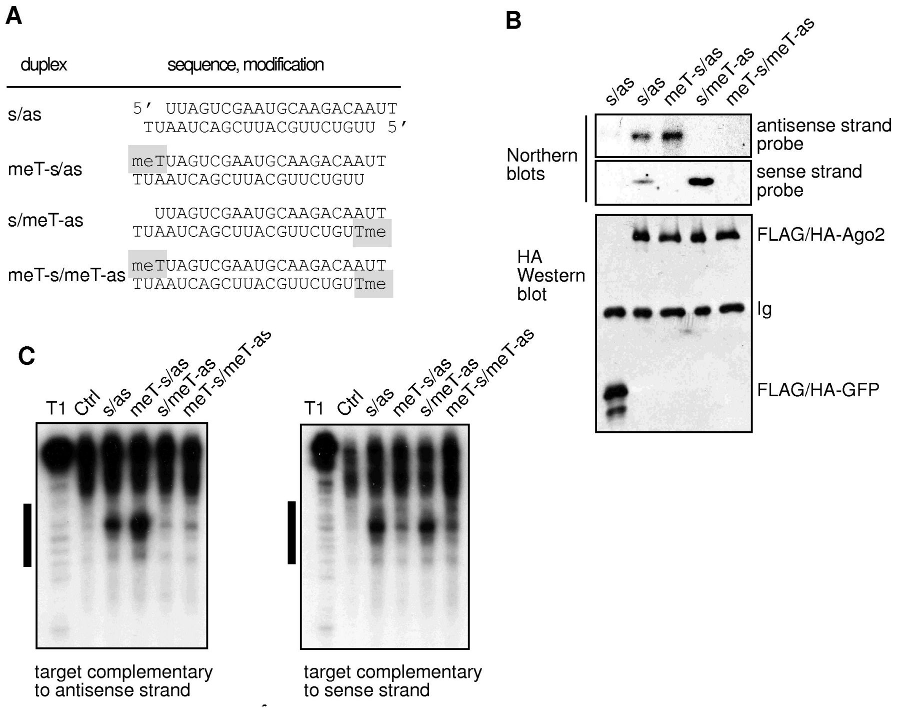


Figure 4.6: siRNA 5'-O-methylation inhibits RISC loading and RISC activity. (A) Schematic presentation of the symmetric RISC loading siRNA duplexes. (B) FLAG/HA-Ago2 and FLAG/HA-EGFP were transiently transfected into HEK 293 cells. Cell lysates were pre–incubated with siRNA duplexes allowing for RISC loading. RISCs were immunoprecipitated using anti–FLAG antibodies and the precipitated proteins were analyzed using anti–HA antibodies (lower panel). Ig indicates the heavy chain of the immunoglobulin. The bound siRNA strands were examined by Northern blotting (upper panel). (C) HeLa cell extracts were pre–incubated with the indicated siRNA duplexes allowing for RISC loading. Control (Ctrl) refers to luciferase siRNA duplex. ^{32}P-cap labeled substrates either complementary to the sense strand or the antisense strand were subsequently added and the cleaved RNA products were analyzed by 4% denaturing RNA PAGE. The bar to the left of the image indicates siRNA target site.

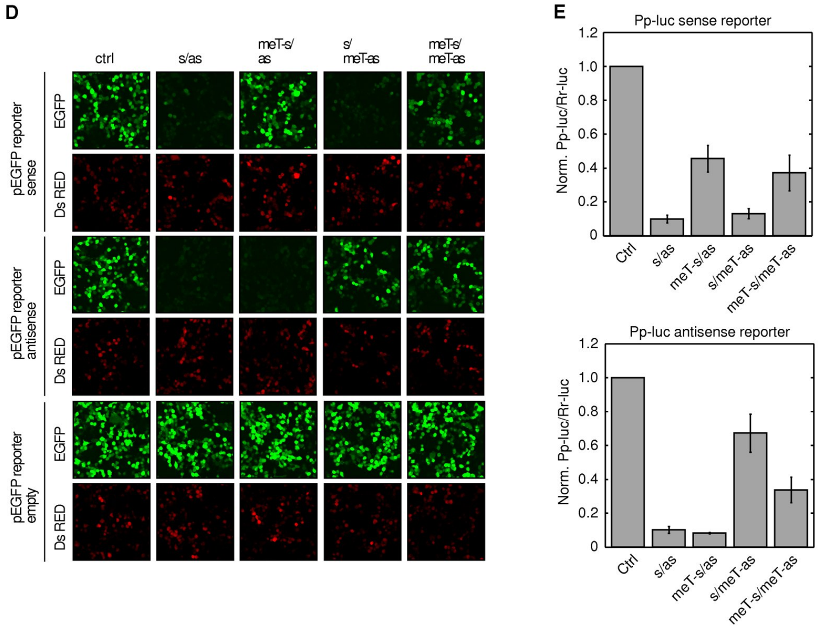


Figure 4.7: siRNA 5'-O-methylation inhibits RISC loading and RISC activity. (D) 5'-O-methylated siRNAs inhibits RNAi in living cells. Plasmids encoding EGFP, EGFP fused to a complementary target site for the antisense strand or EGFP fused to a complementary target site for the sense strand were co–transfected with control (ctrl); luciferase siRNAs or the indicated siRNA duplex. A plasmid encoding the DS Red gene was co–transfected and served as a transfection control. (E) Plasmids containing either a Pp-luciferase gene fused to a complementary target site for the antisense strand or a complementary target site for the sense strand of the siRNA were co–transfected with Rr-luciferase and the indicated siRNAs. GFP siRNA was utilized as control (ctrl) siRNA. The Pp-luc/Rr-luc ratios were normalized to that of the control siRNAs.

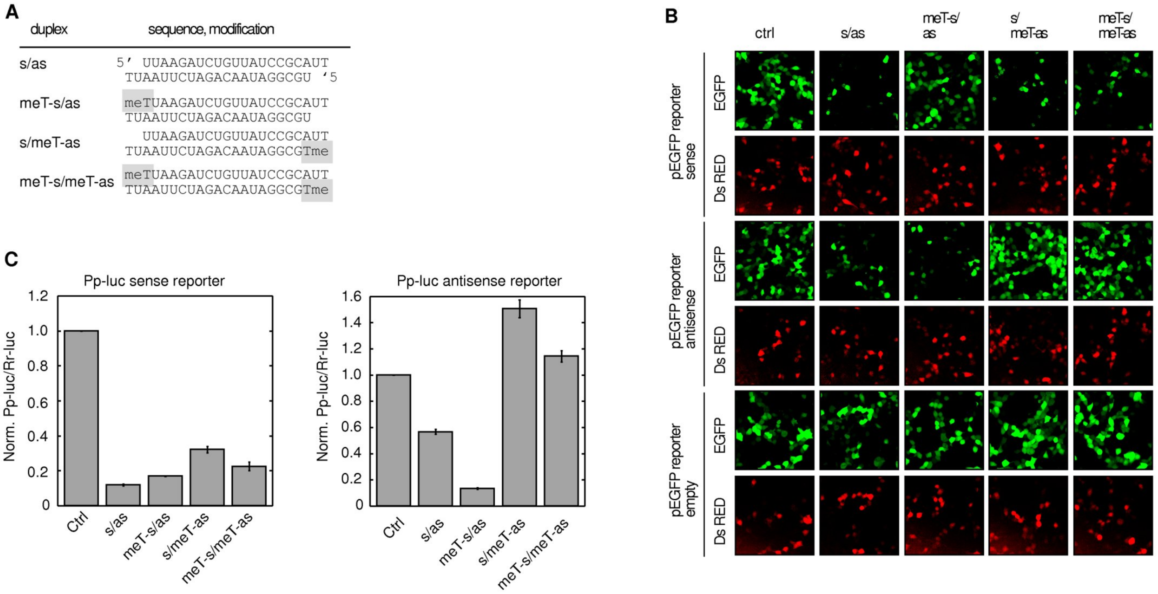


Figure 4.8: siRNA 5'-O-methylation inhibits RISC loading and RISC activity. (A) Schematic presentation of the asymmetric RISC loading siRNA duplexes used in (B) and (C). (B) The same experiments and controls described in (Fig. 2, D) were carried out using the asymmetrically RISC loading siRNAs duplex as well as complementary EGFP target constructs. (D) The same experiments and controls described in (Fig. 2, E) were performed using the asymmetrical RISC loading siRNAs.

4.6A).

The choice of cell line and transfection reagents was critical for identifying siRNA strand– specific off–targets (see Methods). We first tested Lipofectamine 2000 Transfection Reagent (Invitrogen) in HEK 293 cells but variations in gene expression, caused presumably by mild toxicity of the formulated transfection reagent, made it impossible to detect siRNA–sequence dependent off–target signatures. Next, we tested Lipofectamine RNAiMAX Transfection Reagent (Invitrogen) in HEK 293 cells. Though the siRNA and mock transfection yielded reproducible and stable expression profiles, we did not detect any off–targeting signature using either Affymetrix or Agilent mRNA microarrays (data not shown). Finally, we examined HeLa cells transfected with Lipofectamine RNAiMAX using Affymetrix microarrays and we were able to identify the expected off–targeting effects. Using the same approach, we also detected the previously reported off–targets of a siRNA duplex targeting the PIK3CB gene (PIK3CB-6340) [106], indicating that our Affymetrix array platform was sufficiently sensitive for off–target analysis in this cell type(see Methods).

The off–target effects were quantified by analyzing the frequency of seedcomplementary sites (where the seed was defined as nucleotides 1-7, 2-8 or 1-8 of the siRNA) for real and randomized siRNAs within the 3' UTRs of mRNAs that were downregulated one day after siRNA transfection (Figure 4.9). Consistent with our biochemical analysis, which indicated that both strands of our unmodified siRNA duplex were loaded into RISCs, we found seed–complementary site enrichment for both siRNA strands. 5'-O-methylation of the sense strand increased the off–targeting activity of the antisense strand, while decreasing its own off–targeting activity, and vice versa. Note that the sense strand has intrinsically fewer off–targets compared to the antisense strand. This is because the sense strand contains a CG dinucleotide in its seed sequence and matches to CG-containing motifs that are underrepresented in the genome. These analyses thus indicate that chemical modification can limit the off–targeting activity to only one strand of the siRNA duplex.

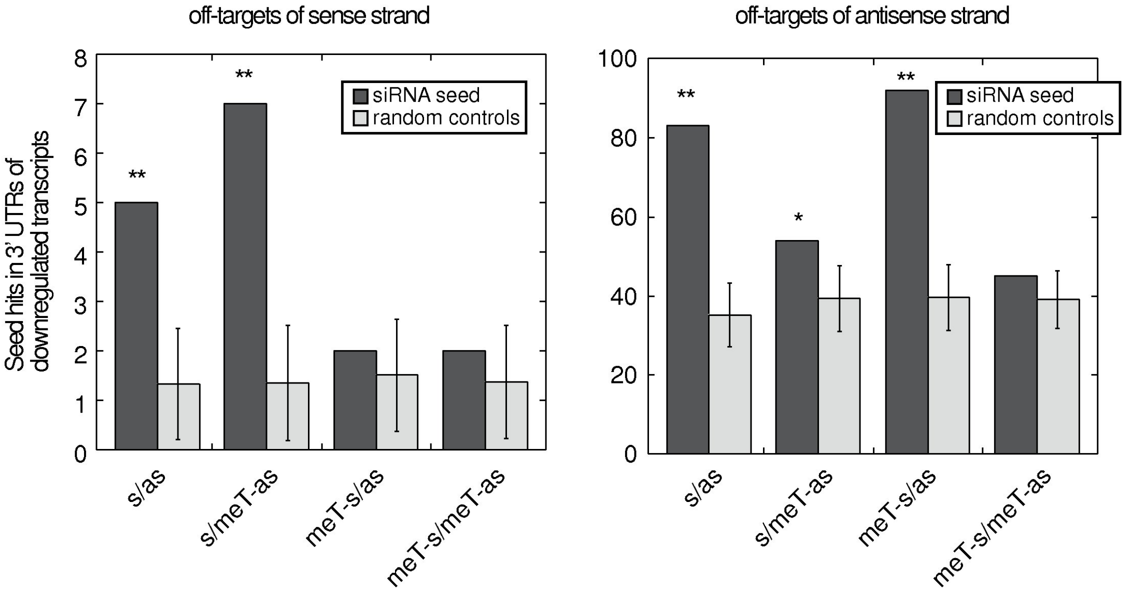


Figure 4.9: siRNA off–target analysis of the symmetric RISC loading siRNA duplexes. Enrichment of seed–complementary sequences for sense (left panel) and antisense (right 34 panel) strands relative to random controls in the 3' UTRs of transcripts that are downregulated upon siRNA transfection (see Methods). Dark grey and light grey bars show the number of occurrences of seed–complementary sites for the siRNAs used in the study, and random controls, respectively. Double and single stars indicate enrichments that are significant at 0.01 and 0.05 level.

4.2.6 Comparison of the effects of 5'-O-methylation and duplex–destabilizing mutations on strand selection

The differential thermodynamic stability of siRNA duplex termini impacts siRNA strand selection [168, 169]. Thermodynamic biases can be introduced by varying the G/C content of the termini of the siRNA duplex or by placing destabilizing, non–Watson–Crick base pairs (mismatches) at one of the termini. We therefore wanted to compare the strand bias introduced by 5'-O-methylation with that introduced by mismatches using the asymmetrical siRNA duplex described above, whose sense strand is preferentially incorporated into RISC.

We placed mismatches in the G/C-rich terminus by altering the sequence of the antisense siRNA from position 1 to 5 (Figure 4.10A). According to the current model for of strand selection, destabilizing the G/C-rich termini should lower the bias for incorporation of the sense strand of this siRNA duplex and should enhance the incorporation of the antisense strand into RISC. What we observed was that mismatches only minimally reduced the activity of the sense siRNA–containing RISC measured by the sense reporter, while 5'-O-methylation of the sense siRNA showed a more pronounced effect (Figure 4.10B).

Antisense siRNA strand incorporation was determined by the antisense reporter assay. Mismatches introduced by altering the sequence of the antisense siRNA trivially lead to mismatches between the antisense siRNA and its reporter, resulting in the lack of cleavage activity for mutants of position 3 to 5 of the antisense siRNA. Mutations placed at position 1 or 2 of the antisense siRNA showed similar activity as the unmodified siRNA duplex, consistent with unaltered behavior of the sense reporter, indicating that mismatches at pos. 1 or 2 were insufficient to alter asymmetry of RISC assembly. In contrast, the 5'-O-methylation of the sense strand, led to a much more pronounced activity of the antisense siRNA. Together, these experiments showed that 5'-O-methylation of the 5' end of siRNAs was more effective in changing strand preferences compared to alterations of thermodynamic stability induced by duplex–destabilizing mismatches.

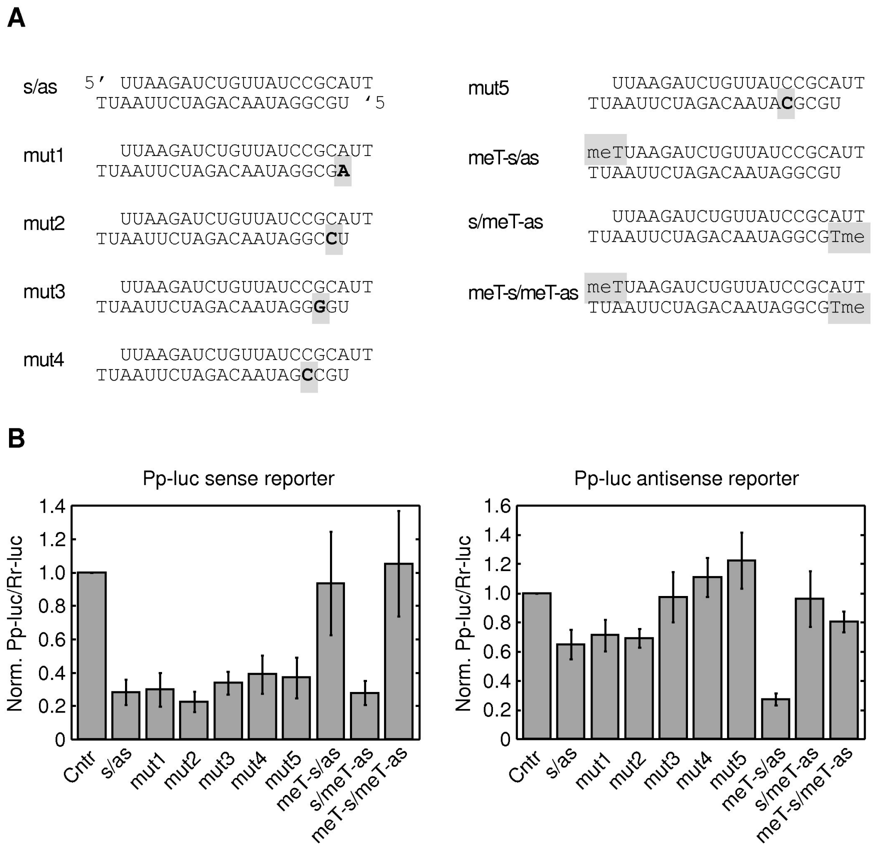


Figure 4.10: Influence of thermodynamic stability vs. 5'-O-methyl modification. (A) Schematic presentation of the asymmetric RISC loading siRNA duplexes used in (B). (B) Plasmids containing either a Pp-luciferase gene fused to a complementary target site for the antisense strand or a complementary target site for the sense strand of the siRNA were co–transfected with Rr-luciferase and the indicated siRNAs. GFP siRNA was utilized as control (Ctrl) siRNA.

4.2.7 Discussion

Our analysis describes for the first time the consequences of strand–specific 5'-O-methylmodification of siRNA duplexes on the assembly and activity of RISC. 5'-O-methylation of the terminal ribose blocked the phosphorylation of siRNAs by cellular kinases [118, 128, 143]. The requirement for 5' phosphates during the assembly of RISC was noted previously using symmetrically 5'-O-methylated siRNA duplexes in D. melanogaster embryo lysates [143], and similarly, we also found that double 5'-O-methyl modification reduced RISC assembly in human cells. Strikingly, the placement of a single 5'-O-methyl only reduced the incorporation of the methylated strand without impairing the incorporation of the unmodified siRNA strand. It even appeared that modification of the sense siRNA strand enhanced the incorporation of the unmodified antisense strand, and vice versa. The effect of the 5'-O-methyl modification was also strong enough to counteract the otherwise strong strand preference of an asymmetrically loading siRNA duplex. This observation emphasizes the importance of 5' phosphate recognition during RISC assembly and its potential use for siRNA design and application.

The molecular events responsible for 5' phosphate recognition during RISC assembly remain to be defined. Presumably the contacts are made while placing the 5' phosphate of the guide siRNA strand into the 5'-phosphate–binding pocket of Argonaute [137, 138]. The sensing of the 5' phosphate also appears to take place in the context of a duplex siRNA or a partially unwound duplex siRNA, because artificial loading of RISC with single–stranded 5'-O-methyl–modified or unmodified siRNAs was possible and led to similar RISC–mediated cleavage activities. Interestingly, a bulky 5' fluorescein modification coupled via a 5' phosphodiester linkage to the guide strand of a siRNA duplex did not affect its silencing efficiency [172]. It will be interesting to explore the effects of other 5'-hydroxyl modifications, for example bulky alkyl groups (e.g. tertiary butyl), on RISC assembly.

Although siRNA duplexes are widely used in research as reagents for gene silencing, sequence–specific off–target effects can be problematic. Off–target effects are typically caused by siRNA sense and antisense strands and the strand–specific reduction of off–target effects using 5'-O-methyl modifications will help to improve siRNA specificity. Although this may also lead to an enhanced off–target effect of the unmodified strand as its loading into RISC may increase, it should nonetheless allow for lowering the dose of siRNAs needed in gene silencing experiments and will aid in controlling potential side effects caused by competition of siRNA with miRNA pathways [173].

Although it has been proposed that thermodynamic asymmetry can be readily introduced into siRNAs using LNAs [174], minimizing the degree of chemical modification needed to control asymmetry may be beneficial from the point of view of manufacturing or potential therapeutic use. It was also interesting to observe that the introduction of conventional mismatches for destabilizing one of the siRNA termini was less effective then the selective placement of a 5'-O-methyl group. In summary, in this study we describe a novel approach of biasing siRNA strand selection from duplex siRNA based on 5' phosphate sensing during RISC assembly.

4.2.8 Materials and Methods

Oligonucleotide synthesis

siRNAs were chemically synthesized using RNA phosphoramidites (Pierce, USA) on an Akta Oligopilot 10 DNA/RNA synthesizer (GE Healthcare Life Sciences) at 1-μmol scale. The synthesis, deprotection and precipitation were performed according to the manufacturer's protocol. 5'-O-methylated siRNAs were purchased from Dharmacon. The sequences of the siRNAs used in this study are as followed: luciferase siRNA antisense strand, 5' UCGAAGUAUUCCGCGUACGUU, sense strand, 5'CGUACGCGGAAUACUUCGAUU; GFP siRNA antisense strand, 5' GGCAAGCUGACCCUGAAGUUT, sense strand, 5'ACUUCAGGGUCAGCUUGCCUT; symmetrically RISC–loaded siRNA antisense strand, 5' UUGUCUUGCAUUCGACUAAUT, sense strand, 5' UUAGUCGAAUGCAAGACAAUT; asymmetrically RISC–load siRNA antisense strand, 5' UUAAGAUCUGUUAUCCGCAUT, sense strand, 5' UGCGGAUAACAGAUCUUAAUT. For analysis of the importance of the 5' phosphate, the 5' uridine residues were substituted by 5'-O-methyl-2'-deoxythymidine or 2'- deoxythymidine as control.

Plasmids

The mammalian expression plasmids for FLAG/HA-tagged Ago2 and GFP were previously described [114] and are available from www.addgene.org. Reporter plasmids for measuring RISC activity of antisense and sense siRNA strands were generated as follows: Complementary pairs of DNA oligonucleotides bearing the siRNA target sequence and flanking SacI and NaeI restriction sites were annealed, digested with SacI and NaeI and cloned into the 3' UTR of the reporter vectors, using the SacI and NaeI sites of the pMIR16 REPORT plasmid (Ambion) or the SacI and SmaI sites of the pEGFP-C2 plasmid (Clontech), respectively. Prior to this procedure, the pEGFP-C2 plasmid was modified by the insertion of a stop codon into the BglII site. The oligos containing the siRNA target sequences were: target complementary to symmetric RISC loading antisense siRNA strand, 5'CGCTGAGCTC ATCGCCACCT TGTTTAAGCC TTAGTCGAAT GCAAGACAAA TTAGACCTAC GCACTCCAGG CCGGCTCGC and 5'GCGAGCCGGC CTGGAGTGCG TAGGTCTAAT TTGTCTTGCA TTCGACTAAG GCTTAAACAA GGTGGCGATG AGCTCAGCG; target complementary to symmetric RISC loading sense siRNA strand, 5'CGCTGAGCTC ATCGCCACCT TGTTTAAGCC TTGTCTTGCA TTCGACTAAA TTAGACCTAC GCACTCCAGG CCGGCTCGC and 5'GCGAGCCGGC CTGGAGTGCG TAGGTCTAAT TTAGTCGAAT GCAAGACAAG GCTTAAACAA GGTGGCGATG AGCTCAGCG; target complementary to asymmetric RISC loading antisense siRNA strand, 5'CGCTGAGCTC ATCGCCACCT TGTTTAAGCC GCGGATAACA GATCTTAAAA TTAGACCTAC GCACTCCAGG CCGGCTCGC and 5'GCGAGCCGGC CTGGAGTGCG TAGGTCTAAT TTTAAGATCT GTTATCCGCG GCTTAAACAA GGTGGCGATG AGCTCAGCG; target complementary to asymmetric RISC loading sense siRNA strand, 5'CGCTGAGCTC ATCGCCACCT TGTTTAAGCC TTTAAGATCT GTTATCCGCA TTAGACCTAC GCACTCCAGG CCGGCTCGC and 5'GCGAGCCGGC CTGGAGTGCG TAGGTCTAAT GCGGATAACA GATCTTAAAG GCTTAAACAA GGTGGCGATG AGCTCAGCG. Plasmids for in vitro transcription of RNA cleavage substrates were generated by cloning of the annealed oligodeoxynucleotides into the SacI/SmaI sites of the pIVEX2.4d plasmid (Roche).

In vitro transcription of RISC cleavage substrates

DNA templates for in vitro transcription of RNA cleavage substrates were generated by linearization of the respective pIVEX2.4d plasmids using BamHI restriction enzyme digestion. The linearized plasmid was used for run–off in vitro transcription using T7 RNA polymerase (Fermentas) according to the manufacturer's protocol. The transcripts were purified by denaturing PAGE, visualized by UV–shadowing, excised and eluted overnight in 0.3 M NaCl at 4°C. The eluted RNA was precipitated by addition of 3 volumes of ethanol and collected by centrifugation. The cleavage substrate complementary to the luciferase targeting siRNAs was ^{32}P cap labeled [165] and the cleavage was assayed as described previously [114]. The sense and antisense ^{32}P cap labeled cleavage substrates were 188 nt and 195 nt, respectively.

Northern blotting

The oligodeoxynucleotide probes 5' TTAGTCGAATGCAAGACAA, and 5' TTGTCTTGCATTCGACTAA were used for detection of the symmetrically RISC–loading siRNAs. Probes were radioactively labeled with T4 polynucleotide kinase (New England Biolabs) using [$\gamma-^{32}$P]-ATP. RNA samples were separated by 15% denaturing PAGE and transferred to a nylon membrane (Hybond-N+, Amersham) by semi–dry electroblotting. The membrane was then subjected to UV crosslinking using the auto–crosslink function on the Stratalinker (Stratagene) and subsequently baked for 1 h at 80°C. The membrane was incubated with 15 ml of pre–hybridization buffer (5xSSC/20 mM sodium phosphate buffer (pH 7.2)/7% SDS/1xDenhardt's solution/3 mg of sonicated salmon sperm DNA) in a hybridization oven for 1 h rotating at 50°C. The pre–hybridization solution was then replaced with 15 ml of hybridization buffer containing 3,000,000 cpm of ^{32}P-radiolabeled DNA probe and incubated overnight at 50°C. The membrane was washed twice with 100 ml of wash solution I (5xSSC/1% SDS) at 50°C for 10 min followed by a single wash step with wash solution II (1xSSC/1% SDS) at 50°C. The membrane was wrapped in plastic wrap and exposed to a film for 2 days.

Tissue culture and transfections

HEK 293 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 100 unit/ml penicillin, and 100 μg/ml streptomycin at 37°C in a 5% CO_2-containing atmosphere. The stably transfected HeLa S3 FLAG/HA-Ago2 cell line [114] was cultured under the same conditions with the addition of 0.5 mg/ml G418. HEK 293 cells were co–transfected with reporter or control plasmids and siRNAs using Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol in a 24–well format 24 h after seeding. Transfected cells were analyzed for luciferase activity or GFP fluorescence 24 h after transfection. For immunoprecipitation experiments HEK 293 cells were transfected using the calcium chloride method. Cells were plated to 40% confluency 3 to 4 h before transfection on a 15 cm dish. 20 μg of plasmid DNA was diluted in 858 μl of water and 122 μl 2M $CaCl_2$. 1 ml of 2x HEPES–buffered saline (274 mM NaCl, 1.5 mM Na_2HPO_4, 54.6 mM HEPES-KOH, pH 7.1) was added drop–wise under gentle agitation. The transfection solution was then sprinkled onto the cells.

Microscopy

HEK 293 cells were seeded on coverslips and were co–transfected with either empty pEGFP-C2 as control or pEGFP-C2 bearing the target complementary to siRNA strand specific sequences (0.2 μg/well), pDsRedmonomer-C1 (Clontech) (0.1 μg/well) and siRNA (20 pmol/well). After 24 h, cells were fixed in PBS with 4% formaldehyde for 30 min at room temperature, washed twice with PBS and mounted to slides using Vectashield mounting medium (Vector Laboratories, Burlingame, CA). Images were recorded using a Leica TCS SP2 confocal laser microscope and a 20x immersion oil objective. For GFP and DsRed images, 10 z–sections of the cells were recorded and processed to average projections using the Leica confocal software.

Dual luciferase assays

For experiments described in Figure 4.5C, 7000 HEK 293 cells per well were transferred into 96-well plates the day before transfection. The cells were then co–transfected with 0.2 μg pGL2-control (Promega), 0.02 μg pRL-TK (Promega), and 3.75 pmol siRNA duplexes (final concentration 25 nM) with Lipofectamine 2000 (Invitrogen, 0.75 μl). Luciferase activities were measured 20 h after transfection using the Dual Luciferase Assay Kit (Promega) and a Bio-Tek Clarity luminometer. The ratios of the signals of firefly (Pp) luciferase to seapansy (Rr) luciferase were calculated and normalized by dividing by the ratio for control siRNA against FLJ30525.3. The plotted data were averaged from triplicates s.d. For Figures 4.6 4.7 4.8 4.9 4.10, HEK 293 cells were cultured in 24 well plates and each co–transfected with either empty pMIR-REPORT (Pp-luc) control plasmid or pMIR-REPORT bearing the target complementary to siRNA strand specific sequences (0.2 μg/well), pRL-SV40 control vector (Rr-luc) (Promega) (0.1 μg/well) and siRNA (20 pmol/well). The cells were lysed and assayed 24 h post transfection following the Dual-Luciferase Reporter Assay system (Promega) instructions. Samples were analyzed on a Mithras LB 940 Multimode Microplate Reader (Berthold Technologies). All samples were assayed in triplicates.

Western blotting, extract preparation and immunoprecipitation

Western blotting was performed as previously described [114]. For immunoprecipitation, HEK 293 cells transiently transfected with FLAG-HA-tagged Ago2 and HeLa S3 cells stably transfected with FLAG/HA-tagged Ago2 were harvested from 15 cm plates 48 h post transfection. The cells were washed with PBS (pH 7.4) and subjected to lysis with 700 μl of lysis buffer (150 mM KCl, 25 mM Tris-HCl (pH 7.4), 2 mM EDTA (pH 8.0), 1 mM NaF, 0.5 mM DTT, 0.05% NP40, 0.5 mM AEBSF) at 4°C for 10 min. The cells were subsequently scraped off of the plate and centrifuged at 17,200 g for 10 min. The supernatant was incubated with 15 μl of anti-FLAG M2 agarose beads (Sigma), which were activated by washing once with 0.1 M glycine-HCl (pH 2.5) and equilibrated by washing with 1.5 M Tris- HCl (pH 8.8), for 3 h at 4°C with rotation. The beads were collected and washed three times with 300 mM NaCl/5 mM $MgCl_2$/0.05% NP40/50 mM Tris-HCl (pH 7.5) and once with PBS (pH 7.5). To isolate the RISC–incorporated siRNA, the beads were then incubated in 300 μl of proteinase K solution consisted of 2x proteinase K buffer (300 mM NaCl/25 mM EDTA (pH 8.0)/2% SDS/200 mM Tris-HCl (pH 7.5) and proteinase K at 1 mg/ml concentration) 37°C for 10 min. The RNA was phenol/chloroform–extracted and ethanol precipitated.

RNA cleavage assays

In vitro transcribed cleavage substrates were 5' cap labeled as described previously [165]. In a typical RNA cleavage reaction 100 nM of siRNA was incubated in a 15 μl reaction containing 50% HeLa S100 extract, 1 mM ATP, 0.2 mM GTP, 10 U/ml RNasin (Promega), 100 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, 10 mM HEPES-KOH (pH 7.9) at 30°C. After 30 min 5 nM of the cap–labeled cleavage substrate was added and further incubated at 30°C for 1.5 h. The reactions were stopped by adding 200 μl proteinase K buffer containing 1 mg/ml proteinase K. The RNA was subsequently isolated using phenol/chloroform extraction and the cleavage products were analyzed by 8% denaturing RNA PAGE. The labeled RNA was detected by phosphoimaging and autoradiography.

Microarray mRNA expression analysis

HeLa cells were plated in a 6-well plate with a volume of 2.5 ml at a density such that next day they are approximately 75% confluenent for transfection. A transfection solution of 0.5 ml was utilized for transfection for the final siRNA duplex concentration of 50 nM using RNAiMAX (Invitrogen). Total RNA was extracted 24 h post transfection with Trizol (Invitrogen) and purified using the RNeasy Mini Kit (Qiagen). 3 μg of purified RNA was used to synthesize the first strand of cDNA using ArrayScript reverse transcriptase (Ambion, Cat #1791) and an oligo(dT) primer bearing a T7 promoter. The single–stranded cDNA was converted into double–stranded DNA (dsDNA) by DNA polymerase I in the presence of E. coli RNase H and DNA ligase. After column purification, the dsDNA was served as a template for in vitro transcription in a reaction containing biotin–labeled UTP, unlabeled NTPs and T7 RNA Polymerase. The amplified, biotin–labeled antisense RNA (aRNA) was purified and quality was assessed using the Agilent 2100 Bioanalyzer and the RNA 6000 Nano kit. Twenty μg of labeled aRNA was fragmented and fifteen μg of the fragmented aRNA was hybridized to Affymetrix Human Genome U133 Plus 2.0 Array for 16 hours at 45°C as described in the Affymetrix Technical Analysis Manual (Affymetrix, Santa Clara, CA). After hybridization, Gene Chips were stained with streptavidin–phycoerythrin, followed by an antibody solution (anti–streptavidin) and a second streptavidin–phycoerythrin solution, with all liquid handling performed by a GeneChip Fluidics Station 450. Gene

Chips were then scanned with the Affymetrix GeneChip Scanner 3000. Agilent Whole Human Genome Oligonucleotide Microarrays (Catalog-Nr. 4112F) were performed by Cogenics.

Computational analyses of putative off–targets

We normalized the probe intensities for the five microarrays (transfection reagent only, symmetric RISC loading sense/antisense, MeO-sense/antisense, sense/MeO-antisense, MeO-sense/ MeO-antisense siRNA duplexes) using the bioconductor (see http://www.bioconductor.org and Gentleman et al., 2004) and gcRMA software (Wu et al., 2004). To quantify off–target effects based on the frequency of seed–complementary sites in the 3' UTRs, we selected, for each gene measured by the microarray, the transcript with median 3' UTR length. This dataset consisted of 14,997 transcripts. From all the probe sets corresponding unambiguously to a given gene, we selected the one that responded best (exhibited the highest variance) across a large number of experiments performed on the Affymetrix platform that we used. This probe set was used to monitor the per–transcript expression level across our experiments. From each experiment (sense/antisense vs. mock transfection, MeO-sense/antisense vs. mock transfection, sense/MeO-antisense vs. mock transfection, MeO-sense/MeO-antisense vs. mock transfection) we extracted the top 1%, i.e. 149, most downregulated transcripts and computed the number of occurrences of matches to the 1-7, 2-8, 1-8 nucleotide positions of both the sense and the antisense strands in these set of transcripts. We compared these numbers with the number of occurrences expected for random siRNAs, which we calculated as follows. For each of the antisense and the sense strand, we determined the number of occurrences of seed–complementary sites in the entire set of 3' UTRs. From all octameric sequences, we then selected the 5% (3277) whose reverse complements occurred with a frequency closest to that of siRNA–complementary sites in the entire set of 3' UTRs. These served as random siRNA controls. We determined the number of occurrences of seed–complementary sites of these random siRNAs in the downregulated set of 3' UTRs, To correct for slight variations in this number that can be expected simply because the frequency of seed–complementary sites for real and random siRNAs in the 3' UTRs overall are not identical, we adjusted the observed counts by a factor equal to the ratio of observed occurrences of the real siRNA–complementary sites to the random siRNAcomplementary sites in the entire set of 3' UTRs. We used the distribution of the values determined for random siRNAs to estimate the p-value of the number of occurrences of real siRNA–complementary sites.

Control experiment:Transfection of the PIK3CB-6340 siRNA

Before starting transfecting siRNAs with chemical modifications we decided to first try to reproduce the transfection experiments for the PIK3CB-6340 siRNA performed by Jackson et. al [106]. We started off by using a HEK293 cell line and the Affymetrix Human Genome U133 Plus 2.0 Array platform. They used an Agilent microarray platform and performed all experiments in a HeLa cell line. Figure 4.11 shows the comparison of the two transfection experiments. We could not detect an agreement between their off–targets and the off–targets determined in our experiment. The PIK3CB gene (main target of the siRNA) which has a perfect complementary site served as a control. And even for the PIK3CB gene we detected substantial differences between the knock down efficiency of our transfection experiment and the experiment performed by them (see Figure 4.11). Taken together, the data show that we could not reproduce the off targets of the PIK3CB-6340 siRNA, probably due to low transfection efficiency or toxicity.

We then changed to a HeLa cell line and to an Agilent array platform which were used in [106] and repeated the experiment. Furthermore we also tested two different transfection reagents, namely oligofectamine and RNAi Max. Figure 4.12 shows MA plots for the two transfection experiments as well as two controls. Both transfection experiments (oligofectamine and RNAi Max) managed to downregulate the main target PIK3CB (see Figure 4.12 E1,E3). RNAi Max showed a slightly stronger potency than oligofectamine. As a first control experiment we compared siRNA transfection with oligofectamine to siRNA transfection with RNAi Max. If it is indeed the case that RNAi Max has a greater potency then we should see an upregulation of the PIK3CB gene in this control experiment. Figure 4.12 E2 shows that this is true. As a second control experiment we compared Transfection control(RNAi Max) to Transfection control(Oligofectamine) which should not affect the expression of PIK3CB. Figure 4.12 E4 shows that this is case. The results indicate that the PIK3CB-6340

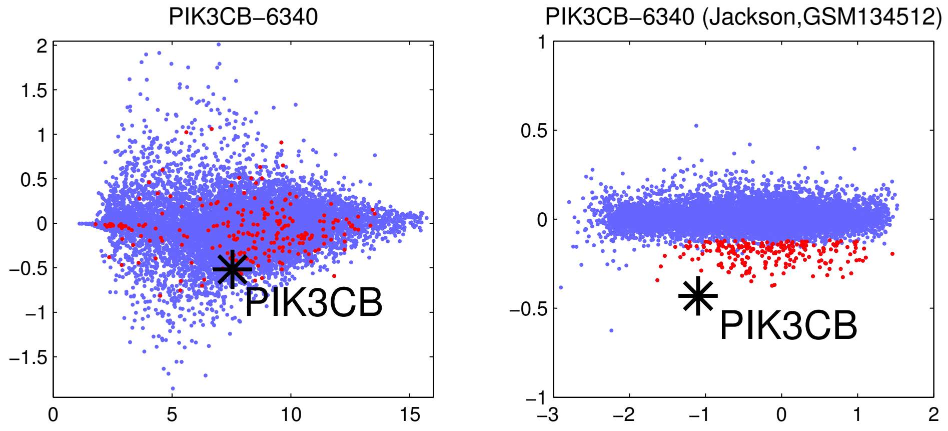


Figure 4.11: PIK3CB-6340 off-target comparison. Right panel shows an MA plot (log expression vs. log fold change) of the PIK3CB transfection experiment (Agilent microarray platform) performed in [106]. Highlighted in red are all the off-targets ($p < 0.01$). The left panel shows the MA plot of our transfection experiment with the off targets from the right panel highlighted in red. The main target PIK3CB is highlighted in black in both panels.

tranfection experiments in HeLa cells worked.

We compared our new Agilent PIK3CB-6340 transfection data (E1-4) to the data from [106]. Figure 4.13 shows the results. We could detect a clear correlation for both oligofectamine(E1) and RNAi Max(E3) transfection reagents (see Figure 4.13A,C). In the case of the control experiment E2 where we compared two siRNA transfections with different reagents we detected the expected anti–correlation (increased potency of RNAi Max compared to oligofectamine). In the case of the control experiment E4(only transfection reagents) we could detect no correlation as we would expect. Taken together the data indicate that not only the transfections worked but that we can also reproduce the off–targets of PIK3CB using HeLa cells and Agilent array technology.

To test whether the array platform (Agilent vs. Affymetrix) could play a crucial role we decided to hybridize the material from the PIK3CB-6340 transfection experiment E3(RNAi Max) to an Affymetrix Human Genome U133 Plus 2.0 array and to compare the results to the reference experiment [106] (see Figure 4.13E) as well as to the data we obtained previously by hybridizing the same material to an Agilent array

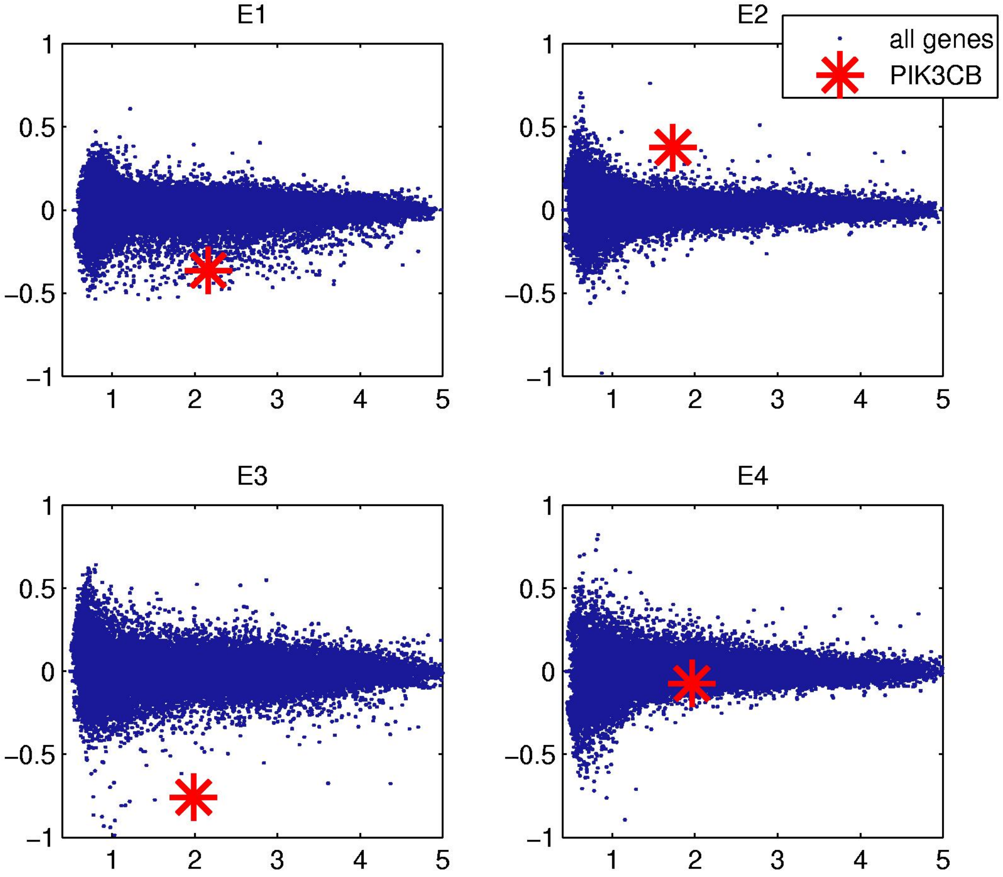


Figure 4.12: PIK3CB-6340 transfection in a HeLa cell line using Agilent array technology. All 4 panels represent MA plos with average expression (log10) on the x-axis and fold change (log10) on the y-axis. E1:Transfection control(oligofectamine) vs. siRNA (oligofectamine). E2: siRNA(RNAi Max) vs. siRNA(Oligofectamine). E3:Transfection control(RNAi Max) vs. siRNA(RNAi Max). E4: Transfection control(RNAi Max) vs. Transfection control(Oligofectamine). The main target PIK3CB is highlighted in red.

(see Figure 4.13F). In both cases we detect a clear correlation indicating that the array platform does not play a crucial role when measuring siRNA off–target activity.

Taken everything together, the data indicate that our siRNA transfection procedure works, that the main target is knocked down, that the off–target activity can be monitored in reproducible fashion and that the results do not depend on the array technology platform(Affymetrix vs. Agilent).

Acknowledgments

We thank S. Rottmüller for technical assistance, P. Linsley, A. Jackson and J. Burchard for guidance on the technical aspects of microarray–based off–targeting analysis, C. Ender and An. Frohn for critically reading the manuscript and S. Jentsch for support. We also acknowledge W. Zhang and C. Zhao for Affymetrix array profiling carried out in the Rockefeller University Genomics Resource Center. L. P. is recipient of a fellowship from the Boehringer–Ingelheim Funds, and D. G. is supported by the Swiss Institute of Bioinformatics. Y.P. was supported by the Ruth L. Kirschstein Fellowship from NIH/National Institute of General Medical Sciences. This work was also supported by the Max Planck Society, HHMI and NIH grant R01 GM068476-01.

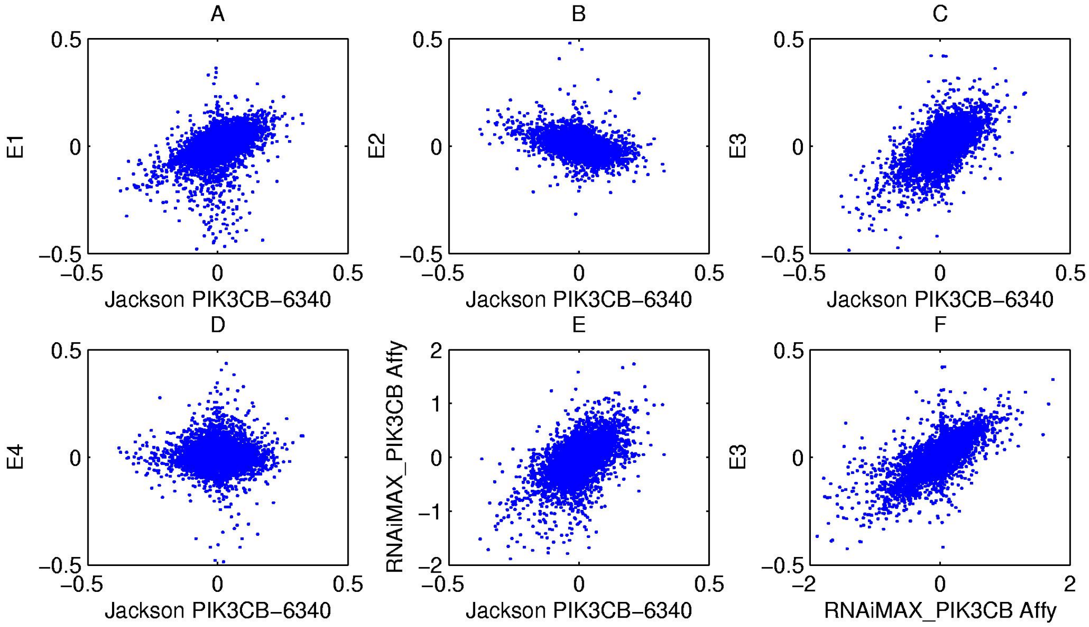


Figure 4.13: PIK3CB-6340 off-target comparison in HeLa cells. Panels A-D show expression fold change scatter plots comparing PIK3CB-6340 from [106] with our PIK3CB-6340 transfection experiments E1-4(Agilent). E1:Transfection control(oligofectamine) vs. siRNA(oligofectamine). E2: siRNA(RNAi Max) vs. siRNA(Oligofectamine). E3:Transfection control(RNAi Max) vs. siRNA(RNAi Max). E4: Transfection control(RNAi Max) vs. Transfection control(Oligofectamine). In Panel E we compared the PIK3CB-6340 transfection experiment from [106] to the RNAi Max experiment (E3) hybridized to an Affymetrix Human Genome U133 Plus 2.0 array. Panel F shows a comparison between the Agilent and the Affy platform using the same material for hybridization (RNAi Max).

Chapter 5

Conclusions

Presented in the previous chapters are the computational tools developed to annotate and characterize small RNA genes and to identify their targets. One of these tools is oligomap, a novel software for fast and exhaustive identification of nearly-perfect matches of small RNAs in sequence databases. In contrast to sequence search programs that rely on heuristics (like e.g. Blast), oligomap is guaranteed to find all the hits with zero or one error (mismatch or indel) to the target sequence. Given a large number of query sequences (in the range between 10^5 and 10^6) oligomap will perform the search one or two orders of magnitude faster than general sequence search tools. This enables fast annotation of large numbers of sequence tags typically generated by deep sequencing technologies. Oligomap is part of an automated annotation pipeline used in our laboratory to annotate small RNA sequences. The application of these tools to samples of small RNAs obtained from mouse and human germ cells together with subsequent computational analyses lead to the discovery of a new class of small RNAs which are now called *piRNAs*. The computational analysis revealed that piRNAs have a strong uridine preference at their 5' end, that unlike miRNAs, piRNAs are not excised from fold-back precursors but rather from long primary transcripts, and that the genome organization of their genes is conserved between human and mouse even though piRNAs on the sequence level are poorly conserved. Piwi proteins, the binding partners of piRNAs are known to be important for stem– and germ–cell development in different animals. Therefore the identification of piRNAs provides an important molecular link regarding the function of Piwi proteins. Recently is has been shown that at least a subpopulation of piRNAs in d.melanogaster are involved in transposon silencing [9] and that an amplification loop exists that could give rise to novel piRNAs once a starting population of piRNAs exists. This suggests the presence of double stranded processing intermediates. Even though we did not observe evidence for this in our initial study, more extensive data obtained by deep sequencing technology [10] revealed a more detailed picture which now also supports the idea of an amplification and a role for piRNAs in transposon silencing in mouse.

In vertebrates, the most studied class of small regulatory RNAs are the miRNAs which bind to mRNAs and block translation. A computational framework is introduced to identify miRNA targets in mammals, flies, worms and fish based on extensive cross species conservation information. It extends the already available methods in several ways. First, it treats the phylogenetic relationships between species in a rigorous and general way, without any tunable parameters. That is, the Bayesian procedure uniquely determines the posterior probabilities for each conservation pattern and seed type in terms of the observed conservation patterns of target sites for each miRNA. The method can be applied to any clade of species, and the phylogenetic relations between the species will be automatically taken into account. Second, the evolution of selection pressures on target sites is estimated in a miRNA-specific manner. This enables the correct treatment miRNAs that appeared at different stages in evolution, and whose targets may have undergone different selection pressures in different lineages. In particular, different miRNAs show markedly different distributions of functional target sites across the phylogenetic tree. This provides the first comprehensive picture of species-specific and clade-specific miRNA targeting. Downstream analysis of the predicted targets has revealed that, especially in long 3' UTRs that occur in vertebrates, miRNA target sites show a significant bias toward occurrence near the start and end of the 3' UTR. The algorithm performs at least as well as the most accurate methods available today, with a high specificity over a relatively large range of sensitivities. Finally to more robustly infer the function of individual miRNAs, each of whom may target hundreds of transcripts, we developed a method for identifying biochemical pathways that are significantly enriched or depleted in targets of a specific miRNA. For well-studied miRNAs, this approach recovers the known functional associations. In addition, this analysis predicts novel pathway associations for a significant number of miRNAs. The result of this study boils down to a long list of predicted miRNA targets. But it's use goes beyond an in silico miRNA target atlas. The data could also be used in the future to learn about the determinants of miRNA targeting. In particular this would mean to locate features which distinguish conserved seed complementary sites from non-conserved ones. As most of the competitor miRNA target prediction methods our method focuses on the seed type binding model and preferential binding in the 3' UTRs of mRNAs. Several lines of evidence suggest that this is the major miRNA targeting paradigm but it's also clear that there are many exceptions to the rule. A genome wide quantification of the contribution of 3'UTR seed type binding compared to non seed type binding or binding in the coding region still remains to be done. We and others have shown that signals for non typical miRNA targeting are rather week but correct quantification of the uncertainties associated with those types of predicted sites could still improve the overall accuracy of the predictions. It has been shown that miRNAs can act by mRNA degradation and/or translational repression. Computational predictions of miRNA targets cannot distinguish between these two possibilities but the intersection with future proteomics data and mRNA degradation data from microrray studies might shed light on the origin of the evolutionary pressure that governs the creation and destruction of miRNA target sites.

SiRNAs have been shown to produce off-targets that resemble miRNA targets. This opens, in principle, the possibility to learn about additional determinants of miRNA target site functionality from siRNA off-target data, which is available in larger amounts. Analysis of published microarray data obtained in siRNA experiments suggests the presence of additional determinants of miRNA target site functionality (beyond complementarity between the miRNA 5' end and the target) in the close vicinity (about 150 nucleotides) of the miRNA-complementary site. Even though the molecular basis of possible determinants is not revealed, the analysis first suggests the presence of additional features which justifies further investment in this topic and second provides additional information about the unknown determinants which can guide the search strategies in the future. We can envision different types of determinants that are compatible with a model of short range dependencies. Local secondary structure, general compositional biases or auxiliary binding motives for either components of the RISC complex itself or other RNA binding proteins that could interact with RISC. Microarray methodology currently represent the only unbiased and genome wide experimental procedure to learn about miRNA binding. But its not clear yet how far these kind of results will generalize. All the experiments that we have analyzed were performed in HeLa cell lines. In the case of auxiliary binding partners of RISC one could imagine target requirements to change across tissues or cell lines. Gaining insight into such processes is fundamental to the understanding of miRNA targeting and we will learn about those processes as soon as the experimental data will become available.

Finally, as part of a study aiming to reduce siRNA off-target effects by introducing chemical modifications in the siRNA, we performed microarray data analysis of siRNA transfection experiments. Presented are the methods used to quantify off-target activity of siRNAs carrying different types of chemical modifications. The analysis revealed that off-targets caused by the passenger strand of the siRNA can be reduced by 5'-O-methylation. Eliminating the incorporation of the passenger strand of the siRNA into RISC does not only eliminate its off-targets, it also causes less perturbation to the endogenous miRNA population in RISC. Taken together, this approach can help in the future to design siRNAs with a higher quality than the ones available today.

Bibliography

[1] P. D. Zamore, T. Tuschl, P. A. Sharp, and D. P. Bartel, "RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals.," Cell **101,** 25–33 (2000).

[2] A. J. Giraldez, Y. Mishima, J. Rihel, R. J. Grocock, S. V. Dongen, K. Inoue, A. J. Enright, and A. F. Schier, "Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs.," Science **312,** 75–79 (2006).

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

[4] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel, "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.," Science **294,** 858–862 (2001).

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

[6] T. A. Volpe, C. Kidner, I. M. Hall, G. Teng, S. I. S. Grewal, and R. A. Martienssen, "Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi.," Science **297,** 1833–1837 (2002).

[7] B. J. Reinhart and D. P. Bartel, "Small RNAs correspond to centromere heterochromatic repeats.," Science **297,** 1831 (2002).

[8] T. Sijen and R. H. A. Plasterk, "Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi.," Nature **426,** 310–314 (2003).

[9] J. Brennecke, A. A. Aravin, A. Stark, M. Dus, M. Kellis, R. Sachidanandam, and G. J. Hannon, “Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila.,” Cell **128,** 1089–1103 (2007).

[10] A. A. Aravin, R. Sachidanandam, A. Girard, K. Fejes-Toth, and G. J. Hannon, “Developmentally regulated piRNA clusters implicate MILI in transposon control.,” Science **316,** 744–747 (2007).

[11] P. H. Olsen and V. Ambros, “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 (1999).

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

[13] S. Pfeffer *et al.*, “Identification of virus-encoded microRNAs.,” Science **304,** 734–736 (2004).

[14] S. Pfeffer *et al.*, “Identification of microRNAs of the herpesvirus family.,” Nat Methods **2,** 269–276 (2005).

[15] P. Landgraf *et al.*, “A mammalian microRNA expression atlas based on small RNA library sequencing.,” Cell **129,** 1401–1414 (2007).

[16] M. Margulies *et al.*, “Genome sequencing in microfabricated high-density picolitre reactors.,” Nature **437,** 376–380 (2005).

[17] S. Bennett, “Solexa Ltd.,” Pharmacogenomics **5,** 433–438 (2004).

[18] A. Aravin *et al.*, “A novel class of small RNAs bind to MILI protein in mouse testes.,” Nature **442,** 203–207 (2006).

[19] H. B. Houbaviy, M. F. Murray, and P. A. Sharp, “Embryonic stem cell-specific MicroRNAs.,” Dev Cell **5,** 351–358 (2003).

[20] G. Meister and T. Tuschl, “Mechanisms of gene silencing by double-stranded RNA.,” Nature **431,** 343–349 (2004).

[21] C. C. Mello and D. Conte, “Revealing the world of RNA interference.,” Nature **431,** 338–342 (2004).

[22] M. A. Matzke and J. A. Birchler, “RNAi-mediated pathways in the nucleus.,” Nat Rev Genet **6,** 24–35 (2005).

[23] P. D. Zamore and B. Haley, “Ribo-gnome: the big world of small RNAs.,” Science **309,** 1519–1524 (2005).

[24] M. A. Carmell, Z. Xuan, M. Q. Zhang, and G. J. Hannon, “The Argonaute family: tentacles that reach into RNAi, developmental control, stem cell maintenance, and tumorigenesis.,” Genes Dev **16,** 2733–2742 (2002).

[25] D. N. Cox, A. Chao, J. Baker, L. Chang, D. Qiao, and H. Lin, “A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal.,” Genes Dev **12,** 3715–3727 (1998).

[26] P. W. Reddien, N. J. Oviedo, J. R. Jennings, J. C. Jenkin, and A. S. Alvarado, “SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells.,” Science **310,** 1327–1330 (2005).

[27] S. Kuramochi-Miyagawa *et al.*, “Mili, a mammalian member of piwi family gene, is essential for spermatogenesis.,” Development **131,** 839–849 (2004).

[28] W. Deng and H. Lin, “miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis.,” Dev Cell **2,** 819–830 (2002).

[29] A. A. Aravin, M. Lagos-Quintana, A. Yalcin, M. Zavolan, D. Marks, B. Snyder, T. Gaasterland, J. Meyer, and T. Tuschl, “The small RNA profile during Drosophila melanogaster development.,” Dev Cell **5,** 337–350 (2003).

[30] D. P. Bartel, “MicroRNAs: genomics, biogenesis, mechanism, and function.,” Cell **116,** 281–297 (2004).

[31] P. Y. Chen *et al.*, “The developmental miRNA profiles of zebrafish as determined by small RNA cloning.,” Genes Dev **19,** 1288–1293 (2005).

[32] A. R. Bellv, J. C. Cavicchia, C. F. Millette, D. A. O’Brien, Y. M. Bhatnagar, and M. Dym, “Spermatogenic cells of the prepuberal mouse. Isolation and morphological characterization.,” J Cell Biol **74,** 68–85 (1977).

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

[34] K. D. Pruitt, T. Tatusova, and D. R. Maglott, “NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.,” Nucleic Acids Res **33,** D501–D504 (2005).

[35] K. Mochizuki and M. A. Gorovsky, “Small RNAs in genome rearrangement in Tetrahymena.,” Curr Opin Genet Dev **14,** 181–187 (2004).

[36] S. I. S. Grewal and J. C. Rice, “Regulation of heterochromatin by histone methylation and small RNAs.,” Curr Opin Cell Biol **16,** 230–238 (2004).

[37] S. Jenab and P. L. Morris, “Testicular leukemia inhibitory factor (LIF) and LIF receptor mediate phosphorylation of signal transducers and activators of transcription (STAT)-3 and STAT-1 and induce c-fos transcription and activator protein-1 activation in rat Sertoli but not germ cells.,” Endocrinology **139,** 1883–1890 (1998).

[38] S. Pfeffer, M. Lagos-Quintana, and T. Tuschl, “Cloning of small RNA molecules,” Current Protocols in Molecular Biology **26,** 4.1–4.18 (2003).

[39] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl, “Identification of tissue-specific microRNAs from mouse.,” Curr Biol **12,** 735–739 (2002).

[40] L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel, “Vertebrate microRNA genes.,” Science **299,** 1540 (2003).

[41] E. Berezikov, V. Guryev, J. van de Belt, E. Wienholds, R. H. A. Plasterk, and E. Cuppen, "Phylogenetic shadowing and computational identification of human microRNA genes.," Cell **120,** 21–24 (2005).

[42] M. Legendre, A. Lambert, and D. Gautheret, "Profile-based detection of microRNA precursors in animal genomes.," Bioinformatics **21,** 841–845 (2005).

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

[44] J. G. Ruby, C. Jan, C. Player, M. J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D. P. Bartel, "Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans.," Cell **127,** 1193–1207 (2006).

[45] A. Girard, R. Sachidanandam, G. J. Hannon, and M. A. Carmell, "A germline-specific class of small RNAs binds mammalian Piwi proteins.," Nature **442,** 199–202 (2006).

[46] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller, "A greedy algorithm for aligning DNA sequences.," J Comput Biol **7,** 203–214 (2000).

[47] E. Berezikov *et al.*, "Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis.," Genome Res **16,** 1289–1298 (2006).

[48] P. Carninci *et al.*, "The transcriptional landscape of the mammalian genome.," Science **309,** 1559–1563 (2005).

[49] D. Gaidatzis, E. van Nimwegen, J. Hausser, and M. Zavolan, "Inference of miRNA targets using evolutionary conservation and pathway analysis.," BMC Bioinformatics **8,** 69 (2007).

[50] "Small RNA group at the Biozentrum Basel[http://www.mirz.unibas.ch/],".

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

[52] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin, "Computational identification of Drosophila microRNA genes.," Genome Biol **4,** R42 (2003).

[53] J. Dostie, Z. Mourelatos, M. Yang, A. Sharma, and G. Dreyfuss, "Numerous microRNPs in neuronal cells containing novel microRNAs.," RNA **9,** 180–186 (2003).

[54] M.-R. Suh *et al.*, "Human embryonic stem cells express a unique set of microRNAs.," Dev Biol **270,** 488–498 (2004).

[55] I. Bentwich *et al.*, "Identification of hundreds of conserved and nonconserved human microRNAs.," Nat Genet **37,** 766–770 (2005).

[56] E. Wienholds, W. P. Kloosterman, E. Miska, E. Alvarez-Saavedra, E. Berezikov, E. de Bruijn, H. R. Horvitz, S. Kauppinen, and R. H. A. Plasterk, "MicroRNA expression in zebrafish embryonic development.," Science **309,** 310–311 (2005).

[57] J. M. Cummins *et al.*, "The colorectal microRNAome.," Proc Natl Acad Sci U S A **103,** 3687–3692 (2006).

[58] H. Xu, X. Wang, Z. Du, and N. Li, "Identification of microRNAs from different tissues of chicken embryo and adult chicken.," FEBS Lett **580,** 3610–3616 (2006).

[59] W. P. Kloosterman, F. A. Steiner, E. Berezikov, E. de Bruijn, J. van de Belt, M. Verheul, E. Cuppen, and R. H. A. Plasterk, "Cloning and expression of new microRNAs from zebrafish.," Nucleic Acids Res **34,** 2558–2569 (2006).

[60] E. Berezikov, F. Thuemmler, L. W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, and R. H. A. Plasterk, "Diversity of microRNAs in human and chimpanzee brain.," Nat Genet **38,** 1375–1377 (2006).

[61] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson, "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs.," Nature **433,** 769–773 (2005).

[62] J. Krtzfeldt, N. Rajewsky, R. Braich, K. G. Rajeev, T. Tuschl, M. Manoharan, and M. Stoffel, "Silencing of microRNAs in vivo with 'antagomirs'.," Nature **438,** 685–689 (2005).

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

[64] D. Grn, Y.-L. Wang, D. Langenberger, K. C. Gunsalus, and N. Rajewsky, "microRNA target predictions across seven Drosophila species and comparison to mammalian targets.," PLoS Comput Biol **1,** e13 (2005).

[65] K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel, "The widespread impact of mammalian MicroRNAs on mRNA repression and evolution.," Science **310,** 1817–1821 (2005).

[66] A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen, "Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution.," Cell **123,** 1133–1146 (2005).

[67] P. Sood, A. Krek, M. Zavolan, G. Macino, and N. Rajewsky, "Cell-type-specific signatures of microRNAs on target mRNA expression.," Proc Natl Acad Sci U S A **103,** 2746–2751 (2006).

[68] J. G. Doench and P. A. Sharp, "Specificity of microRNA target selection in translational repression.," Genes Dev **18,** 504–511 (2004).

[69] J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen, "Principles of microRNA-target recognition.," PLoS Biol **3,** e85 (2005).

[70] E. C. Lai, "Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation.," Nat Genet **30,** 363–364 (2002).

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

[72] A. Krek *et al.*, "Combinatorial microRNA target predictions.," Nat Genet **37,** 495–500 (2005).

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

[74] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, “Human MicroRNA targets.,” PLoS Biol **2,** e363 (2004).

[75] X. Xie, J. Lu, E. J. Kulbokas, T. R. Golub, V. Mootha, K. Lindblad-Toh, E. S. Lander, and M. Kellis, “Systematic discovery of regulatory motifs in human promoters and 3’ UTRs by comparison of several mammals.,” Nature **434,** 338–345 (2005).

[76] S. M. Johnson, H. Grosshans, J. Shingara, M. Byrom, R. Jarvis, A. Cheng, E. Labourier, K. L. Reinert, D. Brown, and F. J. Slack, “RAS is regulated by the let-7 microRNA family.,” Cell **120,** 635–647 (2005).

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

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

[79] M. C. Vella, K. Reinert, and F. J. Slack, “Architecture of a validated microRNA::target interaction.,” Chem Biol **11,** 1619–1623 (2004).

[80] M. J. Benton, *Vertebrate Palaeontology* (Unwin Hyman, 1990).

[81] J. Wittbrodt, A. Shima, and M. Schartl, “Medaka–a model organism from the far East.,” Nat Rev Genet **3,** 53–64 (2002).

[82] M. Z. Michael, S. M. O. Connor, N. G. van Holst Pellekaan, G. P. Young, and R. J. James, “Reduced accumulation of specific microRNAs in colorectal neoplasia.,” Mol Cancer Res **1,** 882–891 (2003).

[83] A. Stark, J. Brennecke, R. B. Russell, and S. M. Cohen, “Identification of Drosophila MicroRNA targets.,” PLoS Biol **1,** E60 (2003).

[84] H. Robins, Y. Li, and R. W. Padgett, “Incorporating structure to predict microRNA targets.,” Proc Natl Acad Sci U S A **102,** 4006–4009 (2005).

[85] C. Burgler and P. M. Macdonald, “Prediction and verification of microRNA targets by MovingTargets, a highly adaptable prediction method.,” BMC Genomics **6,** 88 (2005).

[86] D. R. Hipfner, K. Weigmann, and S. M. Cohen, “The bantam gene regulates Drosophila growth.,” Genetics **161,** 1527–1537 (2002).

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

[88] N. S. Sokol and V. Ambros, “Mesodermally expressed Drosophila microRNA-1 is regulated by Twist and is required in muscles during larval growth.,” Genes Dev **19,** 2343–2354 (2005).

[89] C.-Z. Chen, L. Li, H. F. Lodish, and D. P. Bartel, “MicroRNAs modulate hematopoietic lineage differentiation.,” Science **303,** 83–86 (2004).

[90] I. Naguibneva, M. Ameyar-Zazoua, A. Polesskaya, S. Ait-Si-Ali, R. Groisman, M. Souidi, S. Cuvellier, and A. Harel-Bellan, “The microRNA miR-181 targets the homeobox protein Hox-A11 during mammalian myoblast differentiation.,” Nat Cell Biol **8,** 278–284 (2006).

[91] D. J.-F. de Quervain and A. Papassotiropoulos, “Identification of a genetic cluster influencing memory performance and hippocampal activity in humans.,” Proc Natl Acad Sci U S A **103,** 4270–4274 (2006).

[92] J. D. Bui, S. Calbo, K. Hayden-Martinez, L. P. Kane, P. Gardner, and S. M. Hedrick, “A role for CaMKII in T cell memory.,” Cell **100,** 457–467 (2000).

[93] P. T. Nelson, D. A. Baldwin, W. P. Kloosterman, S. Kauppinen, R. H. A. Plasterk, and Z. Mourelatos, “RAKE and LNA-ISH reveal microRNA expression and localization in archival human brain.,” RNA **12,** 187–191 (2006).

[94] A. J. Giraldez, R. M. Cinalli, M. E. Glasner, A. J. Enright, J. M. Thomson, S. Baskerville, S. M. Hammond, D. P. Bartel, and A. F. Schier, “MicroRNAs regulate brain morphogenesis in zebrafish.,” Science **308,** 833–838 (2005).

[95] M. Yanagisawa, K. Nakashima, K. Takeda, W. Ochiai, T. Takizawa, M. Ueno, M. Takizawa, H. Shibuya, and T. Taga, "Inhibition of BMP2-induced, TAK1 kinase-mediated neurite outgrowth by Smad6 and Smad7.," Genes Cells **6,** 1091–1099 (2001).

[96] A. Cimmino *et al.*, "miR-15 and miR-16 induce apoptosis by targeting BCL2.," Proc Natl Acad Sci U S A **102,** 13944–13949 (2005).

[97] Q. Jing, S. Huang, S. Guth, T. Zarubin, A. Motoyama, J. Chen, F. D. Padova, S.-C. Lin, H. Gram, and J. Han, "Involvement of microRNA in AU-rich element-mediated mRNA instability.," Cell **120,** 623–634 (2005).

[98] N. C. Barbet, U. Schneider, S. B. Helliwell, I. Stansfield, M. F. Tuite, and M. N. Hall, "TOR controls translation initiation and early G1 progression in yeast.," Mol Biol Cell **7,** 25–42 (1996).

[99] G. A. Calin *et al.*, "Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.," Proc Natl Acad Sci U S A **99,** 15524–15529 (2002).

[100] N. Zanesi *et al.*, "Effect of rapamycin on mouse chronic lymphocytic leukemia and the development of nonhematopoietic malignancies in Emu-TCL1 transgenic mice.," Cancer Res **66,** 915–920 (2006).

[101] E. van Nimwegen, N. Paul, R. Sheridan, and M. Zavolan, "SPA: a probabilistic algorithm for spliced alignment.," PLoS Genet **2,** e24 (2006).

[102] "Genome Bioinformatics group at the University of California, Santa Cruz [http://genome.cse.ucsc.edu],".

[103] M. Blanchette *et al.*, "Aligning multiple genomic sequences with the threaded blockset aligner.," Genome Res **14,** 708–715 (2004).

[104] A. L. Jackson, S. R. Bartz, J. Schelter, S. V. Kobayashi, J. Burchard, M. Mao, B. Li, G. Cavet, and P. S. Linsley, "Expression profiling reveals off-target gene regulation by RNAi.," Nat Biotechnol **21,** 635–637 (2003).

[105] A. Birmingham *et al.*, “3’ UTR seed matches, but not overall identity, are associated with RNAi off-targets,” Nature Methods **3,** 199–204 (2006).

[106] A. L. Jackson, J. Burchard, J. Schelter, B. N. Chau, M. Cleary, L. Lim, and P. S. Linsley, “Widespread siRNA ”off-target” transcript silencing mediated by seed region sequence complementarity.,” RNA **12,** 1179–1187 (2006).

[107] A. L. Jackson *et al.*, “Position-specific chemical modification of siRNAs reduces ”off-target” transcript silencing.,” RNA **12,** 1197–1205 (2006).

[108] D. S. Schwarz, H. Ding, L. Kennington, J. T. Moore, J. Schelter, J. Burchard, P. S. Linsley, N. Aronin, Z. Xu, and P. D. Zamore, “Designing siRNA that distinguish between genes that differ by a single nucleotide.,” PLoS Genet **2,** e140 (2006).

[109] Y. Dorsett and T. Tuschl, “siRNAs: applications in functional genomics and potential as therapeutics.,” Nat Rev Drug Discov **3,** 318–329 (2004).

[110] C. J. Echeverri and N. Perrimon, “High-throughput RNAi screening in cultured cells: a user’s guide.,” Nat Rev Genet **7,** 373–384 (2006).

[111] F. Fuchs and M. Boutros, “Cellular phenotyping by RNAi.,” Brief Funct Genomic Proteomic **5,** 52–56 (2006).

[112] D. E. Root, N. Hacohen, W. C. Hahn, E. S. Lander, and D. M. Sabatini, “Genome-scale loss-of-function screening with a lentiviral RNAi library.,” Nat Methods **3,** 715–719 (2006).

[113] E. Krausz, “High-content siRNA screening.,” Mol Biosyst **3,** 232–240 (2007).

[114] G. Meister, M. Landthaler, A. Patkaniowska, Y. Dorsett, G. Teng, and T. Tuschl, “Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs.,” Mol Cell **15,** 185–197 (2004).

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

[116] Y. Tomari and P. D. Zamore, "Perspective: machines for RNAi.," Genes Dev **19,** 517–529 (2005).

[117] S. M. Elbashir, W. Lendeckel, and T. Tuschl, "RNA interference is mediated by 21- and 22-nucleotide RNAs.," Genes Dev **15,** 188–200 (2001).

[118] S. Weitzer and J. Martinez, "The human RNA kinase hClp1 is active on 3' transfer RNA exons and short interfering RNAs.," Nature **447,** 222–226 (2007).

[119] H. A. Ebhardt, E. P. Thi, M.-B. Wang, and P. J. Unrau, "Extensive 3' modification of plant small RNAs is modulated by helper component-proteinase expression.," Proc Natl Acad Sci U S A **102,** 13398–13403 (2005).

[120] B. Yu, Z. Yang, J. Li, S. Minakhina, M. Yang, R. W. Padgett, R. Steward, and X. Chen, "Methylation as a crucial step in plant microRNA biogenesis.," Science **307,** 932–935 (2005).

[121] V. V. Vagin, A. Sigova, C. Li, H. Seitz, V. Gvozdev, and P. D. Zamore, "A distinct small RNA pathway silences selfish genetic elements in the germline.," Science **313,** 320–324 (2006).

[122] M. D. Horwich, C. Li, C. Matranga, V. Vagin, G. Farley, P. Wang, and P. D. Zamore, "The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC.," Curr Biol **17,** 1265–1272 (2007).

[123] Y. Kirino and Z. Mourelatos, "Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini.," Nat Struct Mol Biol **14,** 347–348 (2007).

[124] T. Ohara, Y. Sakaguchi, T. Suzuki, H. Ueda, K. Miyauchi, and T. Suzuki, "The 3' termini of mouse Piwi-interacting RNAs are 2'-O-methylated.," Nat Struct Mol Biol **14,** 349–350 (2007).

[125] A. Plisson, E. Sarot, G. Payen-Groschne, and A. Bucheton, "A novel repeat-associated small interfering RNA-mediated silencing pathway downregulates complementary sense gypsy transcripts in somatic cells of the Drosophila ovary.," J Virol **81,** 1951–1960 (2007).

[126] K. Saito, Y. Sakaguchi, T. Suzuki, T. Suzuki, H. Siomi, and M. C. Siomi, "Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi-interacting RNAs at their 3' ends.," Genes Dev **21,** 1603–1608 (2007).

[127] G. Hutvgner and P. D. Zamore, "A microRNA in a multiple-turnover RNAi enzyme complex.," Science **297,** 2056–2060 (2002).

[128] L. A. Martinez, I. Naguibneva, H. Lehrmann, A. Vervisch, T. Tchnio, G. Lozano, and A. Harel-Bellan, "Synthetic small inhibiting RNAs: efficient tools to inactivate oncogenic mutations and restore p53 pathways.," Proc Natl Acad Sci U S A **99,** 14849–14854 (2002).

[129] S. M. Hammond, S. Boettcher, A. A. Caudy, R. Kobayashi, and G. J. Hannon, "Argonaute2, a link between genetic and biochemical analyses of RNAi.," Science **293,** 1146–1150 (2001).

[130] L. Peters and G. Meister, "Argonaute proteins: mediators of RNA silencing.," Mol Cell **26,** 611–623 (2007).

[131] J.-J. Song, J. Liu, N. H. Tolia, J. Schneiderman, S. K. Smith, R. A. Martienssen, G. J. Hannon, and L. Joshua-Tor, "The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes.," Nat Struct Biol **10,** 1026–1032 (2003).

[132] K. S. Yan, S. Yan, A. Farooq, A. Han, L. Zeng, and M.-M. Zhou, "Structure and conserved RNA binding of the PAZ domain.," Nature **426,** 468–474 (2003).

[133] A. Lingel, B. Simon, E. Izaurralde, and M. Sattler, "Nucleic acid 3'-end recognition by the Argonaute2 PAZ domain.," Nat Struct Mol Biol **11,** 576–577 (2004).

[134] J.-B. Ma, K. Ye, and D. J. Patel, "Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain.," Nature **429,** 318–322 (2004).

[135] J. S. Parker, S. M. Roe, and D. Barford, "Crystal structure of a PIWI protein suggests mechanisms for siRNA recognition and slicer activity.," EMBO J **23,** 4727–4737 (2004).

[136] J.-J. Song, S. K. Smith, G. J. Hannon, and L. Joshua-Tor, “Crystal structure of Argonaute and its implications for RISC slicer activity.,” Science **305,** 1434–1437 (2004).

[137] J.-B. Ma, Y.-R. Yuan, G. Meister, Y. Pei, T. Tuschl, and D. J. Patel, “Structural basis for 5’-end-specific recognition of guide RNA by the A. fulgidus Piwi protein.,” Nature **434,** 666–670 (2005).

[138] J. S. Parker, S. M. Roe, and D. Barford, “Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex.,” Nature **434,** 663–666 (2005).

[139] F. V. Rivas, N. H. Tolia, J.-J. Song, J. P. Aragon, J. Liu, G. J. Hannon, and L. Joshua-Tor, “Purified Argonaute2 and an siRNA form recombinant human RISC.,” Nat Struct Mol Biol **12,** 340–349 (2005).

[140] Y.-R. Yuan, Y. Pei, J.-B. Ma, V. Kuryavyi, M. Zhadina, G. Meister, H.-Y. Chen, Z. Dauter, T. Tuschl, and D. J. Patel, “Crystal structure of A. aeolicus argonaute, a site-specific DNA-guided endoribonuclease, provides insights into RISC-mediated mRNA cleavage.,” Mol Cell **19,** 405–419 (2005).

[141] J.-J. Song and L. Joshua-Tor, “Argonaute and RNA–getting into the groove.,” Curr Opin Struct Biol **16,** 5–11 (2006).

[142] Y.-R. Yuan, Y. Pei, H.-Y. Chen, T. Tuschl, and D. J. Patel, “A potential protein-RNA recognition event along the RISC-loading pathway from the structure of A. aeolicus Argonaute with externally bound siRNA.,” Structure **14,** 1557–1565 (2006).

[143] A. Nyknen, B. Haley, and P. D. Zamore, “ATP requirements and small interfering RNA structure in the RNA interference pathway.,” Cell **107,** 309–321 (2001).

[144] Q. Liu, T. A. Rand, S. Kalidas, F. Du, H.-E. Kim, D. P. Smith, and X. Wang, “R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway.,” Science **301,** 1921–1925 (2003).

[145] X. Liu, F. Jiang, S. Kalidas, D. Smith, and Q. Liu, “Dicer-2 and R2D2 coordinately bind siRNA to promote assembly of the siRISC complexes.,” RNA **12,** 1514–1520 (2006).

[146] K. Frstemann, Y. Tomari, T. Du, V. V. Vagin, A. M. Denli, D. P. Bratu, C. Klattenhoff, W. E. Theurkauf, and P. D. Zamore, “Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein.,” PLoS Biol **3,** e236 (2005).

[147] F. Jiang, X. Ye, X. Liu, L. Fincher, D. McKearin, and Q. Liu, “Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila.,” Genes Dev **19,** 1674–1679 (2005).

[148] K. Saito, A. Ishizuka, H. Siomi, and M. C. Siomi, “Processing of pre-microRNAs by the Dicer-1-Loquacious complex in Drosophila cells.,” PLoS Biol **3,** e235 (2005).

[149] Y. Tomari, C. Matranga, B. Haley, N. Martinez, and P. D. Zamore, “A protein sensor for siRNA asymmetry.,” Science **306,** 1377–1380 (2004).

[150] J. B. Preall and E. J. Sontheimer, “RNAi: RISC gets loaded.,” Cell **123,** 543–545 (2005).

[151] T. P. Chendrimada, R. I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura, and R. Shiekhattar, “TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing.,” Nature **436,** 740–744 (2005).

[152] R. I. Gregory, T. P. Chendrimada, N. Cooch, and R. Shiekhattar, “Human RISC couples microRNA biogenesis and posttranscriptional gene silencing.,” Cell **123,** 631–640 (2005).

[153] A. D. Haase, L. Jaskiewicz, H. Zhang, S. Lain, R. Sack, A. Gatignol, and W. Filipowicz, “TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing.,” EMBO Rep **6,** 961–967 (2005).

[154] E. Maniataki and Z. Mourelatos, "A human, ATP-independent, RISC assembly machine fueled by pre-miRNA.," Genes Dev **19,** 2979–2990 (2005).

[155] Y. Lee, I. Hur, S.-Y. Park, Y.-K. Kim, M. R. Suh, and V. N. Kim, "The role of PACT in the RNA silencing pathway.," EMBO J **25,** 522–532 (2006).

[156] C. Matranga, Y. Tomari, C. Shin, D. P. Bartel, and P. D. Zamore, "Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes.," Cell **123,** 607–620 (2005).

[157] T. A. Rand, S. Petersen, F. Du, and X. Wang, "Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation.," Cell **123,** 621–629 (2005).

[158] P. J. F. Leuschner, S. L. Ameres, S. Kueng, and J. Martinez, "Cleavage of the siRNA passenger strand during RISC assembly in human cells.," EMBO Rep **7,** 314–320 (2006).

[159] J. Liu, M. A. Carmell, F. V. Rivas, C. G. Marsden, J. M. Thomson, J.-J. Song, S. M. Hammond, L. Joshua-Tor, and G. J. Hannon, "Argonaute2 is the catalytic engine of mammalian RNAi.," Science **305,** 1437–1441 (2004).

[160] T. A. Rand, K. Ginalski, N. V. Grishin, and X. Wang, "Biochemical identification of Argonaute 2 as the sole protein required for RNA-induced silencing complex activity.," Proc Natl Acad Sci U S A **101,** 14385–14389 (2004).

[161] K. Miyoshi, H. Tsukumo, T. Nagami, H. Siomi, and M. C. Siomi, "Slicer function of Drosophila Argonautes and its involvement in RISC formation.," Genes Dev **19,** 2837–2848 (2005).

[162] Y. Tomari, T. Du, B. Haley, D. S. Schwarz, R. Bennett, H. A. Cook, B. S. Koppetsch, W. E. Theurkauf, and P. D. Zamore, "RISC assembly defects in the Drosophila RNAi mutant armitage.," Cell **116,** 831–841 (2004).

[163] G. Meister, M. Landthaler, L. Peters, P. Y. Chen, H. Urlaub, R. Lhrmann, and T. Tuschl, "Identification of novel argonaute-associated proteins.," Curr Biol **15,** 2149–2155 (2005).

[164] G. B. Robb and T. M. Rana, "RNA helicase A interacts with RISC in human cells and functions in RISC loading.," Mol Cell **26,** 523–537 (2007).

[165] J. Martinez, A. Patkaniowska, H. Urlaub, R. Lhrmann, and T. Tuschl, "Single-stranded antisense siRNAs guide target RNA cleavage in RNAi.," Cell **110,** 563–574 (2002).

[166] X. Lin, X. Ruan, M. G. Anderson, J. A. McDowell, P. E. Kroeger, S. W. Fesik, and Y. Shen, "siRNA-mediated off-target gene silencing triggered by a 7 nt complementation.," Nucleic Acids Res **33,** 4527–4535 (2005).

[167] P. S. Linsley *et al.*, "Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression.," Mol Cell Biol **27,** 2240–2252 (2007).

[168] A. Khvorova, A. Reynolds, and S. D. Jayasena, "Functional siRNAs and miRNAs exhibit strand bias.," Cell **115,** 209–216 (2003).

[169] D. S. Schwarz, G. Hutvgner, T. Du, Z. Xu, N. Aronin, and P. D. Zamore, "Asymmetry in the assembly of the RNAi enzyme complex.," Cell **115,** 199–208 (2003).

[170] Y. Pei and T. Tuschl, "On the art of identifying effective and specific siRNAs.," Nat Methods **3,** 670–676 (2006).

[171] T. Tuschl, P. D. Zamore, R. Lehmann, D. P. Bartel, and P. A. Sharp, "Targeted mRNA degradation by double-stranded RNA in vitro.," Genes Dev **13,** 3191–3197 (1999).

[172] J. Harborth, S. M. Elbashir, K. Vandenburgh, H. Manninga, S. A. Scaringe, K. Weber, and T. Tuschl, "Sequence, chemical, and structural variation of small interfering RNAs and short hairpin RNAs and the effect on mammalian gene silencing.," Antisense Nucleic Acid Drug Dev **13,** 83–105 (2003).

[173] D. Grimm, K. L. Streetz, C. L. Jopling, T. A. Storm, K. Pandey, C. R. Davis, P. Marion, F. Salazar, and M. A. Kay, "Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways.," Nature **441,** 537–541 (2006).

[174] J. Elmn *et al.*, “Locked nucleic acid (LNA) mediated improvements in siRNA stability and functionality.,” Nucleic Acids Res **33,** 439–447 (2005).