Computational methods for analyzing small RNAs and their interaction partners with large-scale techniques

Inauguraldissertation

zur Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

VON
PHILIPP FRIEDRICH BERNINGER
AUS ESCHAU, DEUTSCHLAND
BASEL,2011

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Prof. Dr. Mihaela Zavolan und Prof. Dr. Witold Filipowicz

Basel, den 8.Dezember 2009 Prof. Dr. E. Parlow

Contents

Acknowledgments

I especially want to thank my thesis advisor, Prof. Mihaela Zavolan, for her guidance, encouragement and excellent advices. In addition, I would like to thank my lab colleagues and collaborators as well as my friends and family who have supported and inspired me during my doctoral study.

Biologists should not deceive themselves with the thought that some new class of biological molecules, of comparable importance to the proteins, remains to be discovered. This seems highly unlikely.

Francis Crick, On Protein Synthesis, 1958

Chapter 1

Introduction

1.1 Discovery of small RNA mediated silencing pathways

In the last decade, our understanding of gene regulation in eukaryotes has changed dramatically with the awareness of small RNA mediated gene silencing pathways. The key components of these pathways are $\sim$ 20-30 nucleotides (nt) long single-stranded RNAs and members of the Argonaute protein family, to which small RNAs are bound. The small RNAs guide the Argonaute proteins to their targets, which are identified through (partial) complementarity to the guiding small RNAs. The pathways are present in all eukaryotes, apart from lineage-specific loss, for example in budding yeast. Historically, the first members of RNAi pathways have been discovered in the early 90s by independent groups in different organisms. These findings converged on the discovery of various small RNA mediated silencing pathways.

In 1998, Fire and Mello discovered to their surprise, that exogenous double-stranded RNA (dsRNA) was substantially more effective in gene silencing in *Caenorhabiditis elegans* than sense or antisense strand alone [1]. This mechanism was named RNA interference (RNAi) and because only a few molecules of dsRNA were needed for effective gene silencing. Fire and Mello proposed a catalytic amplification component in the interference process. For their discovery that small interfering RNA (siRNA) causes suppression of gene activity in a homology-dependent manner, Fire and Mellow were awarded the 2006 Nobel Prize for Physiology or Medicine.

In 1993, the first microRNA (miRNA) lin-4 was discovered as a regulator of lin-14 in the nematode *Caenorhabiditis elegans* [2]. Quite surprisingly, the study found that the lin-4 gene did not encode a protein, but a small RNA which showed antisense complementarity to several sites in the 3' UTR of lin-14 [2, 3]. It took another seven years, until the second miRNA, let-7 was identified in *Caenorhabiditis elegans* [4].

Once it was realized that let-7 is strongly conserved in evolution [5], the hunt for additional small regulatory RNAs began, and soon after the list of known miRNAs has enormously expanded [6–8].

The parallelism between the siRNA and miRNA pathway was revealed shortly after, when siRNAs of similar length as miRNAs and common pathway components such as Dicer and Argonaute proteins were discovered in plants and animals [9–13]. This was followed by the identification of various small RNA pathways in plants, fungi and animals. Todate, we are aware of small RNA classes such as miRNAs, endogenous siRNAs (endo-siRNAs) and Piwi-interacting RNAs (piRNAs), which have the ability to regulate a broad variety of biological processes.

1.2 Argonaute proteins

The first characterized member of the Argonaute family proteins was the *Drosophila* Piwi gene (P-element induced wimpy testis), which was identified for its importance in germline stem cell division [14]. Shortly after Ago1 and Zwille were identified as important regulators of plant development in *Arabidopsis thaliana* [15, 16]. Initially, Argonautes were named after the squid-like leafs obtained from a Ago1 mutant in *Arabidopsis thaliana* [15].

Currently, the Argonaute proteins are classified into three different clades [17]. The Argonaute family members or the Argonaute-like proteins are defined by their similarity to the *Arabidopsis thaliana* Ago1 protein. Poteins with high similarity to the Piwi protein in *Drosophila melanogaster* are named Piwi-like proteins or the Piwi family members. The third clade consists of worm specific Argonautes (WAGOs) or group 3 Argonautes. Indicative of their early evolutionary origin and regulatory importance,Argonautes are present in most eukaryotes. Their number ranges from one in *Schizosaccharomyces pombe* up to 26 in *Caenorhabiditis elegans*, however several eukaryotes such as *Saccharomyces cerevisiae* appear to have lost the Argonaute genes.

The Argonaute-like proteins, which are involved in siRNA and miRNA pathways, tend to be ubiquitously expressed in multicellular organisms and associate with $\sim$ 21-23 nt long RNAs derived from both endogenous or exogenous source. The animal-specific Piwi-like proteins maintain genome integrity through the piRNA and scnRNA pathways. They are preferentially expressed in the germ line and bind $\sim$ 24-30 nt long RNAs. The worm specific Argonautes are involved in secondary siRNA pathways.

Argonaute proteins adopt a bilobal structure (reviewed in [18–20]). The fist lobe consist of the N-terminal domain which is required for protein-protein interactions and the PAZ domain, which binds to the 3' end of the small RNA. The second lobe consists of the MID domain, which binds to the 5' end of the small RNA, and the PIWI domain, which catalyzes cleavage of target transcripts in several Argonaute members.

The N-terminal domain is required for protein-protein interactions of Piwi family

members. In *Drosophila melanogaster*, HP1a has been identified to interact directly with Piwi by binding to the N-terminal domain [21]. Quite recently, several members of tudor domain containing proteins have been identified as interaction partners of Piwi proteins, which require the methylation of arginine residues in the N-terminal domain [22–26].

The PAZ domain, which is named after the three first identified Argonaute members Piwi, Argonaute and Zwille, is a RNA-binding domain which exhibits an OB (oligonucleotide/oligosaccharide binding) fold and has the ability to bind to the 3' end of either single-stranded RNA or the 3' end overhang of double-stranded RNA in a sequence-independent manner.

The MID domain, which is centered between the PAZ and the PIWI, contains a pocket-like structure similar to the sugar-binding domain of the Lac repressor. This pocket recognizes the characteristic 5' phosphate and binds the 5' end of the small RNA [27,28]. In structural studies of archeal ternary guide-target-Argonaute complexes, the base of the 5' end is flipped and unpaired [29], which provides structural support to the observation that the nucleotides 2-8 of the guide strand are of major importance for target recognition [30,31]. Apart from RNA-binding, the MID domain is also involved in protein-protein interactions by recognizing a conserved motif termed 'Ago hook' in members of the GW182 family [32].

Structural studies of the C-terminal PIWI domain revealed a strong homology to an RNase H-like fold. RNase H enzymes are endonucleases with a conserved DDH motif which cleave RNA in a RNA-DNA duplex. Although not present in all Argonaute proteins, most of them share a more degenerated DD(H/D/E/K) motif [18] and biochemical studies demonstrated the ability of several Argonaute proteins to cleave target RNA opposite to the 10^{th} and 11^{th} position of the guide RNA, which results in a 3' fragment carrying a 5'-phosphate and a 5' fragment with a 3'-OH end. Argonaute proteins which are endonucleolytically active are also called 'slicers'. In *Drosophila melanogaster*, all five members of the Argonaute family exhibit slicer activity [33–36], while in humans, slicer activity has been shown for Ago2 [37,38]. Interestingly, most Group 3 Argonaute proteins lack the catalytic residues, which suggests a cleavage-independent regulatory activity [17]. Apart from being responsible for the endonucleolytic activity of Argonaute proteins, the PIWI domain contains the PIWI box, which interacts with the RNase III domains of human Dicer [39].

1.3 The miRNA pathway in animals

1.3.1 Biogenesis of miRNAs

MicroRNAs (miRNAs) are $\sim$ 22 nt long endogenous single-stranded RNAs which derive from hairpins formed by short inverted repeats. They act as guides for Argonaute

proteins and identify their targets through Watson-Crick basepairing with their target transcripts. MiRNAs are present in plants, animals and viruses [40], however their biogenesis, the target recognition of miRNAs as well as their regulatory function are quite diverse.

In metazoa, prototypical miRNAs arise from long primary transcripts (pri-miRNAs) or from introns from mRNAs which can give rise to several miRNAs [41]. Those transcripts contain inverted repeat sequences, which fold back to form stem-loop structures. These are recognized by the nuclear microprocessor complex, which consists of Drosha and DGCR8/Pasha [42–45]. DGCR8/Pasha, a dsRNA binding protein, directly interacts with the stem of the pri-miRNA and serves as an anchor for Drosha [46], an RNase III enzyme, which excises the hairpin. The Drosha cut leads to a $\sim$ 70 nt miRNA precursor (pre-miRNA), with the RNAase III characteristic overhang of 2 nt at the 3' end and a 5' monophosphate. Additionally, several miRNAs in introns, called mirtrons, exist, which bypass the Drosha processing step [47–49]. In this case the 3' overhang is a direct result of the splicing process. The pre-miRNA is subsequently transported from the nucleus to the cytoplasm by Ran-GTP and Exportin-5 [50, 51].

In the cytoplasm, another RNase III enzyme, Dicer, recognizes the miRNA precursor stem-loop and generates a $\sim$ 18-24 nt long dsRNA by cleaving off the loop of the hairpin [52]. In mammals, Dicer associates with TRBP and/or PACT, and in *Drosophila melanogaster* Dicer-1 interacts with Loquacious [53–57]. After unwinding the dsRNA duplex, one strand called mature miRNA or guide strand, is incorporated into an Argonaute protein, whereas the other strand, called passenger strand, is degraded [58]. The strand selection depends on the thermodynamic stability of the 5' ends. Generally, the strand with the least stable structure at the 5' end is incorporated into the Argonaute proteins [59, 60]. The association of GW182 or its orthologs with the miRNA-bound Argonaute protein is both necessary and sufficient for the formation of the miRNA-induced silencing complex (miRISC) [61–63].

1.3.2 Target recognition

So far, several hundreds of miRNAs have been identified across the animal kingdom [40]. Computational studies suggested that a large fraction of the transcriptome is regulated by miRNAs [64, 65]. Although rare cases of miRNA target sites with extensive complementarity between miRNA and target exist (and direct Argonaute-catalyzed cleavage [66]), target recognition by partial complementarity is more common in metazoa. The nucleotides 2-8 of miRNA, called miRNA 'seed', guide the miRISC to its targets by Watson-Crick basepairing (reviewed in [67]). The target sites are preferentially located in an AU-rich environment in the 3' UTR of the target mRNA towards the beginning or the end of long 3' UTRs [68, 69]. In case of several target sites within one transcript, they normally act independently [70]. Nonetheless target sites located

within short distance to each other appear to have some sort of synergetic effect [69]. Apart from canonical miRNA target sites, involving perfect basepairing with mRNA 'seed', atypical binding modes have been reported, where binding of the 3' end of the miRNA is also involved in targeting [4, 66]. In those cases, a pairing of at least 3 nt with the positions 13-17 of the miRNA either supplements the 'seed' pairing, or even compensates for a mismatch or a bulge inside the 'seed' region.

1.3.3 Regulatory functions of miRNAs

MiRNAs are implicated in a broad variety of biological processes in development [2, 4, 71–73], cell cycle [74], metabolism [75–77] and diseases [78, 79]. The inhibitory action of miRNAs is achieved through mRNA destabilization and/or translational repression. Though the precise molecular mechanisms are still debated. Most studies indicate that repression occurs during the initiation step by preventing cap-binding, however also several reports argue in favor of post-transcriptional mechanism like inhibition of elongation, premature translation termination through ribosome drop-off or co-translational degradation of proteins (discussed in [62, 80]. The contradictory findings, of how miRNAs inhibit translation, might be a result of experimental shortcomings or miRNA indeed inhibit translation by multiple mechanisms.

Although initial studies suggested that miRNAs inhibit translation without affecting the stability of their target mRNAs, the miRNA destabilization effect is now well established [72,81,82]. Argonaute-directed cleavage is not responsible for target degradation, but rather the recruitment of deadenylation and decapping enzymes [72, 82]. Interestingly, deadenylation still occurs when transcription is globally blocked [83]. Recent large-scale studies identified several miRNA targets that were only translationally repressed [84, 85]. This suggests that mRNA degradation might not be a cause of translation inhibition and vice versa.

1.4 Endo-siRNAs in animals

1.4.1 Introduction to distinct types of endo-siRNAs

Initially, the siRNA pathway was viewed as a defense mechanism of the cell directly against selfish and invasive RNA elements. This was supported by the observations that disruption of the siRNA pathways in *Drosophila melanogaster* and *Caenorhabditis elegans* did not have obvious phenotypes, except for higher susceptibility to viral infections and up-regulation of repetitive elements [86–90]. Today a broad variety of endo-siRNAs are known, such as hairpin-siRNAs (hp-siRNAs) derived from hairpins formed by inverted repeat structures, natural antisense siRNAs (nat-siRNAs) derived from dsRNA formed by complementary transcripts or secondary siRNAs generated by

RNA-dependent RNA polymerases (RdRP). Nat-siRNAs have been further classified as cis-nat-siRNA if sense and antisense transcript are derived from either bidirectional or antisense transcription of the same locus, or as trans-nat-siRNAs, if the two transcripts originate from different genomic loci. Until recently, endo-siRNAs were well characterized in plants [91] and fungi [92, 93] but not in animals.

1.4.2 Biogenesis of endo-siRNAs

In *Caenorhabditis elegans*, the targets of (exogenous) siRNAs act as templates for the production of secondary (endogenous) siRNAs. Targeting a transcript with an exogenous siRNA leads to the recruitment of an RdRP, which synthesizes secondary siRNAs [94, 95]. The secondary siRNAs are antisense to the target transcript and carry a triphosphate at their 5' end, which suggests a Dicer-independent biogenesis [94–96]. The targeted transcript itself acts as template for the RdRP, which produces secondary siRNAs, which subsequently associate with the worm-specific Argonaute members [17, 94, 95].

However, both *Drosophila melanogaster* and mouse lack RdRPs and the initially repeat-associated siRNAs (rasiRNAs) turned out to be a subclass of piRNAs [97]. In *Drosophila melanogaster*, so far identified endo-siRNAs originate from inter- or intramolecular stem structures [98–103]. They associate with the exogenous RNAi pathway members Dicer-2 and Ago2, however, unlike in the exogenous RNAi pathway, some endo-siRNAs in *Drosophila melanogaster* require the interaction of Dicer-2 with an isoform of Loquacious instead of R2D2 [99, 102, 104, 105]. The endo-siRNAs showed an enrichment for both repetitive regions as well as for 3' UTRs [98–103].

In mammals, cells normally do not tolerate long dsRNA fragments [10], but during during early development, the interferon pathway is suppressed in oocytes and embryonic stem cells, and endo-siRNAs have been discovered in murine oocytes [106]. By utilizing deep sequencing, a broad spectrum of endo-siRNAs such as cis-nat-siRNAs, trans-nat-siRNAs and hp-siRNAs have been identified [107, 108]. Quite similar to endo-siRNAs in fly, a high fraction of endo-siRNAs derived from repetitive loci as well as from mRNAs. The dsRNAs giving rise to nat-siRNAs were formed by pairs of mRNA transcripts and pseudogenes. In contrast to oocytes, hp-siRNAs are the predominant endo-siRNAs in embryonic stem cells, and pseudogene derived siRNAs have not been found [109]. However, the abundance of endo-siRNAs seem to be quite low in embryonic stem cells [109–111].

1.4.3 Regulatory functions of endo-siRNAs

The production of secondary siRNAs is the reason why RNAi is so effective in *Caenorhabditis elegans* even with a very small amount of primary siRNAs.

In contrast, the generation of endo-siRNAs in *Drosophila melanogaster* and mouse is not triggered by exogenous siRNAs. Endo-siRNAs in *Drosophila melanogaster* are not restricted to the germ line, and the origin of some of the endo-siRNAs from repetitive loci points towards their function in silencing repetitive elements. Similarly, overlap of endo-siRNAs with piRNA loci in mouse oocytes has been detected. Not all endo-siRNAs in mouse and fly are derived from repetitive loci, some are derived from mRNAs. In fly, the transcripts giving rise to cis-nat-siRNAs enriched preferentially in RNA- and DNA-binding proteins, and mammalian endo-siRNA targets were related to microtubule dynamics, suggesting a more general regulatory role of endo-siRNAs apart from repeat silencing [98–103, 107, 108].

1.5 The piRNA pathway

1.5.1 Introduction to the piRNA pathway

The first evidence of a repeat silencing pathway in *Drosophila melanogaster* was discovered in 2001 [112], but it took five more years to establish a Piwi protein dependent small RNA pathway. Primary characteristic of the 'prototypical' piRNA pathway in metazoa is the expression of both Piwi proteins and Piwi-interacting RNAs (piRNAs) in the germ line or in germ line supporting tissues. Those piRNAs normally cluster to distinct genomic loci. Their 5' ends show a strong preference for uridine and their 3' end carries 2' oxygen methylation. Generally, they are between 23 and 32 nt long and are generated in a Dicer-independent manner. So far, piRNA pathways have been identified in a broad range of animals, such as mammals [106, 113–118], zebrafish [119, 120], clawfrog [121, 122], fruit fly [35, 123, 124], silkworm [125, 126], nematode [96, 127, 128] and flatworm [129, 130].

1.5.2 Piwi proteins and their interaction partners

The founding member of the Piwi family, the *Drosophila melanogaster* Piwi gene, was initially identified in a mutagenesis screen in germ line for genes involved in the asymmetric stem cell division [14]. In Piwi mutants, not only the maintenance of both male and female germ line stem cells is disrupted [14], but also decreased germ cell formation and defects in oogenesis have been observed [131, 132]. The second member of the *Drosophila melanogaster* Piwi family members, Aubergine (Aub), is essential for proper germ cell formation, and the offspring of female Aub mutants lack germ cells [133, 134]. Absence of the third Piwi member, Argonaute3 (Ago3) in *Drosophila melanogaster* results in sterile females and males show defects in germ cell maintenance [135].

The mouse Piwi family consists of three members, Miwi, Mili and Miwi2. They

are predominantly expressed in the male germ line and are essential for spermatogenesis and germ line development. Loss of each individual family member results in infertile males with decreased testes size, whereas female mice do not show any obvious phenotype [136–139]. For Mili, the broadest expression profile has been observed, ranging from arrested germ stem cells in the embryo from 12.5 days post-coitum (dpc) to the round spermatids [22, 138, 140]. Miwi2 is expressed in a short time window, ranging from 15.5 dpc up to three days after birth in arrested germ stem cells [141]. In contrast to the other two family members, Miwi is only expressed in the adult animal from meiotic spermatocytes up to the elongating spermatids stage [137]. The loss of both Mili and Miwi2 was accompanied by spermatogenic stem cell arrest, loss of DNA methylation marks in retrotransposons and increased expression of repetitive elements [141,142], whereas the knockout of Miwi results in a block at the early spermatid stage [137].

In *Caenorhabditis elegans*, disruption of the Piwi protein Prg-1 and Prg-2 results in fertilization defects [127, 128]. The loss of Prg-1 is accompanied by dramatical reduction of germ cells and temperature-dependent fertility defects [128].

Several tudor domain containing proteins have been identified recently to interact with Piwi proteins in mouse and *Drosophila melanogaster* [22–24, 26]. Methylation of arginines in the N-terminal domain is required for this interaction, and is carried out by the arginine methyltransferase PRMT5 in both fruit fly and mouse [24, 26], Responsible for the 3' end 2'-O methylation of piRNAs [119, 143, 144] in fruit fly and mouse is Hen1 [145–147]. Through mutant screens in *Drosophila melanogaster*, several putative piRNA pathway components have been identified in the nuage, such as the Krimper [148], the helicases Armitage [149] and Spindle-E [148], the nucleases Squash, Zucchini [150], Maelstrom [151] and Cutoff [152].

1.5.3 Biogenesis of piRNAs

Early studies of RNA mediated silencing pathways in male testes from *Drosophila melanogaster* revealed the involvement of Aub and $\sim$ 24-29 nt long repeat-associated siRNAs (rasiRNAs) in the silencing of the repetitive Stellate locus [112, 123, 153]. By immunoprecipitation of Piwi complexes, the direct interaction between rasiRNAs and Piwi proteins could be shown [35,124]. Therefore they were classified as piRNAs, with rasiRNAs as a subclass derived from repetitive loci. The majority of piRNAs identified in *Drosophila melanogaster* map to a small set of discrete genomic loci, ranging from several up to hundreds of kilobases in length [97]. These piRNA clusters are located in the heterochromatin and highly enriched in repeat-rich regions [154]. The *flamenco* locus has been identified as a transposon regulatory locus, long before the discovery of piRNAs [155]. Loss of the *flamenco* locus, which consists of a mix of transposable elements, results in an increased expression of retrotransposons, defects in germ cell

development and sterility [134, 156, 157]. Mutations in the 5' end of the *flamenco* locus disrupts the generation of piRNAs downstream of the locus, which suggests that mature piRNAs are generated through random excision of mature piRNAs from a long precursor transcript [134]. The excision of piRNAs from the precursor transcript is Dicer-independent [124], and the mature piRNAs are subsequendly modified at the 3' end [36, 124]. In contrast to Ago3 or Aub expression, the Piwi protein is not restricted to germ cells and can be detected in the nuclei of both somatic and germ cells. The presence of a somatic Piwi-dependent piRNA pathway has been reported recently [135, 158]. Ago3 and Aub are localized in the nuage, an amorphous structure that surrounds the nucleus in germ cells.

In mouse, piRNAs consist of two different populations, according to the developmental stage in which they are expressed. The class of piRNAs that is expressed before the meiotic pachytene is classified as pre-pachytene piRNAs, and is similar to the piRNAs in *Drosophila melanogaster*. This piRNA population associates with both Mili and Miwi2 and originates from genomic loci enriched in retrotransposons when compared to the piRNA loci from the pachytene stage [154]. The sequence pool of pre-pachytene piRNAs is dynamic. Although the fraction of transposon-derived piRNAs is more or less stable during development, the composition of individual transposon classes changes during developmental progression [141]. Comparison of the piRNAs bound to Mili and Miwi2 reveals that both proteins associate with transposon-derived piRNAs, but Miwi2 shows a stronger preference for repeat-derived piRNAs than Mili [141]. The second class of piRNAs, called pachytene piRNAs, are derived from genomic loci with little overlap to pre-pachytene piRNA clusters [154]. They are expressed during the pachytene stage until the haploid round spermatid stage and interact with both Mili and Miwi [154]. A much higher fraction of pachytene piRNAs maps uniquely to the genome compared to the pre-pachytene piRNAs [154].

The piRNAs of *Caenorhabditis elegans* were initially called 21U-RNAs, due to their uracil bias at the 5' end and their length of 21 nt [96]. They carry a 5' monophosphate and a modified oxygen at their 3' end [96]. The majority originates from two large loci of chromosome IV, and has a consensus motif, located $\sim$60 nt upstream of the mature piRNA [96]. Although the upstream motif is conserved between *C. elegans* and *C. briggsae* [96, 128], the piRNA sequences are not conserved and the expression of the individual piRNAs correlates with their match to the consensus motif [128]. It has been hypothesized, that the piRNAs are transcribed individually [96], but the upstream motif might also act as a processing signal on posttranscriptional level [159]. RNAi screens against components of the RNAi pathway components showed depletion of piRNAs in Prg-1, but not in Prg-2 or Dcr-1 mutants [127, 128, 159]. The 21U-RNAs were identified as piRNAs by immunoprecipitation of Prg-1 and by sequencing the associated small RNAs [127, 128, 159].

1.5.4 Regulatory functions of piRNAs

It has been observed that the mature piRNA sequences themselves are poorly conserved between closely related species, even though they are derived from syntenic regions [113, 116, 128]. A strong positive selection and cluster acquisition has been shown for the genomic loci of murine piRNAs, which suggests an evolutionary arms race [160], consistent with the regulatory role. Immunoprecipitation and sequencing of the piRNAs associated with individual Piwi family members revealed distinct species of piRNAs. The piRNAs associated with Piwi and Aub showed a strong bias towards an U at their 5' end and derived preferentially from the antisense-strand of retrotransposons [36, 134]. In contrast to that, piRNAs associated with Ago3 derived from the sense strand of retrotransposons showed a strong bias for an A at the 10^{th} position [36, 134]. Interestingly, the 5' ends of piRNAs associated with either Piwi and Aub overlapped significantly with the 5' ends of Ago3-bound piRNAs from the opposite strand by 10 nt [36, 134]. This observation together with the finding that all three Piwi proteins in *Drosophila melanogaster* are able to cleave target transcripts between the 10^{th} and 11^{th} nt from the 5'end of the small RNA guide strand [36, 145], lead to the proposal of a piRNA amplification loop [36, 134], called 'ping-ping' mechanism.

The amplification loop is initiated by a primary mature piRNA bound to Ago3 that triggers cleavage of an antisense piRNA precursor transcript. This event defines the 5' end of the secondary piRNA, which eventually associates with either Piwi or Aub. The 3' end of the secondary piRNA is assumed to be generated by either an exo- or endonuclease and modified at the 3' end, resulting in a 24-30 nt secondary piRNA. The generation of another piRNA, which is triggered by the secondary piRNA in a similar manner, completes the amplification loop. Distinct size preferences of piRNAs bound to individual Piwi proteins reflect the footprint of the bound protein complex, which protects the tail of the piRNA from the nuclease event [134].

Once the amplification loop is started, both primary and secondary piRNAs are generated, providing the organism an adaptive immune system towards transposable elements. What is not clear at the moment is how the initial species of piRNAs is generated. Some solution to this problem is provided by observations that at least some piRNAs in *Drosophila* are maternally deposited [161], and piRNAs have also been reported in mouse oocytes [107, 108].

Consistent with their repetitive origin, the signatures of the ping-pong mechanism are also found in pre-pachytene piRNAs in mouse [154]. So far, it is assumed that pachytene piRNAs consist only of primary piRNAs, and ping-pong signatures have not yet been found [97, 141, 162]. The loci of pachytene piRNA are depleted of repetitive elements when compared with the average repeat density of the mouse genome, and derepression of transposons has not been reported for Miwi mutants [154]. It has been shown that both Mili and Miwi2 act upstream of DNA methylation pathways in mouse

[141, 142], indicating that transposons are also silenced on transcriptional level in a piRNA-dependent manner. Nonetheless because both Miwi and Mili interact with the cap-binding complex, they probably have a general role in translational control apart from transposon silencing [140, 163].

In general, piRNAs in *Caenorhabditis elegans* do not show sequence similarity to transposons, and so far, no evidence for the ping-pong model has been observed. Their regulatory targets have not yet been identified [127, 128, 159].

Chapter 2

Computational analysis of small RNA cloning data

Philipp Berninger*, Dimos Gaidatzis*, Erik van Nimwegen, Mihaela Zavolan
Division of Bioinformatics, Biozentrum, University of Basel, Switzerland
These authors contributed equally.

published in *Methods 44 , 13–21, 2008*

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. In this report we present the methods that we developed for the annotation and expression profiling of small RNAs obtained through large-scale sequencing. These include a fast algorithm for finding nearly perfect matches of small RNAs in sequence databases, a web-accessible software system for the annotation of small RNA libraries, and a Bayesian method for comparing small RNA expression across samples.

2.1 Introduction

Though recently discovered, small RNAs appear to play a wealth of regulatory roles, ranging from degradation of target mRNA [72, 164], translation silencing of target mRNA [6–8], chromatin remodeling [93,165] and transposon silencing [134,154,166]. In vertebrates, the most studied class of small regulatory RNAs are the microRNAs (miRNAs), which are produced from hairpin precursors by the Dicer endonuclease [6–8] to block the translation of target mRNAs [167]. The discovery of the let-7 miRNA, which is perfectly conserved in sequence from worm to man [4], sparked a great interest in the identification of additional miRNAs as well as of other regulatory RNAs. The group of Tom Tuschl developed a protocol for isolating miRNAs which typically yields $80-90\%$ miRNAs in a given sample of small RNAs [168, 169], and used it to collect small RNA expression profiles from hundreds of mammalian samples. Based on this data, we constructed an atlas of miRNA expression profiles in a large number of mammalian tissues [170]. In parallel, high-throughput pyrosequencing [171] or sequencing-by-synthesis [172] 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. Here we present the methods that we have developed and used to identify novel regulatory RNAs and to analyze their expression across cells and tissues [113, 170].

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

2.2.1 Problem definition

Figure 2.1 shows a sketch of the protocol for small RNA sequencing. 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 [169]. 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. We accomplish this by mapping the adaptors to the concatamer sequences using WU-BLAST (http://blast.wustl.edu). Because the rate sequencing errors is rather high (0.01-0.03/nucleotide), and the 5'/3' adaptors short (10-16 nucleotides), we use a set of parameters that permit the identification of short, imperfect matches between query and target sequences. These pa-

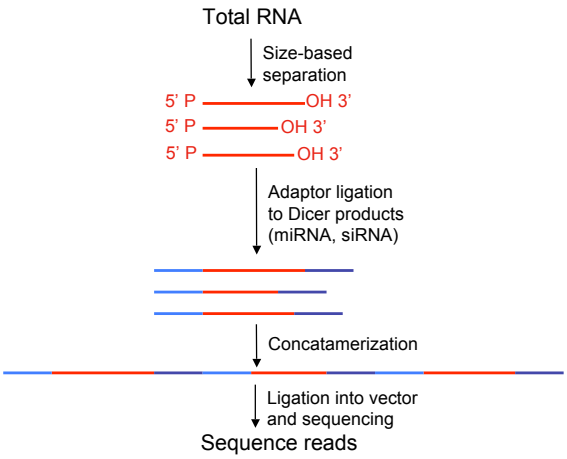


Figure 2.1: Protocol for small RNA sequencing.

rameters are: short initial matches between adaptor and cloned sequence (W = 4), low score thresholds determining which of the initial alignments are to be extended (S2=50 gapS2=50), and relatively large drop in the score determining when an alignment is not to be further extended (X=50 gapX=50). The subsequences of a concatamer that are found between matches to 5' and 3' adaptors in the correct configuration are extracted as small RNAs.

Functional 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 [173], such as WU-BLAST (http://blast.wustl.edu) [170], Blast [96,114], or Megablast ([174]) [175] 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. While this was not a problem when we needed to map a few adaptors to concatamers, it became a problem when we tried 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.

2.2.2 Oligomap algorithm

A sketch of the main components of the algorithm is shown in Figure 2.2. The approach is to build a tree from the input small RNA sequences (Figure 2.2C) and then search this tree with subsequences starting at each position of the target sequence (Figure 2.2D). 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 2.2A). The search stage is performed through a number of "walkers" (Figure 2.2B). 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.

2.2.3 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

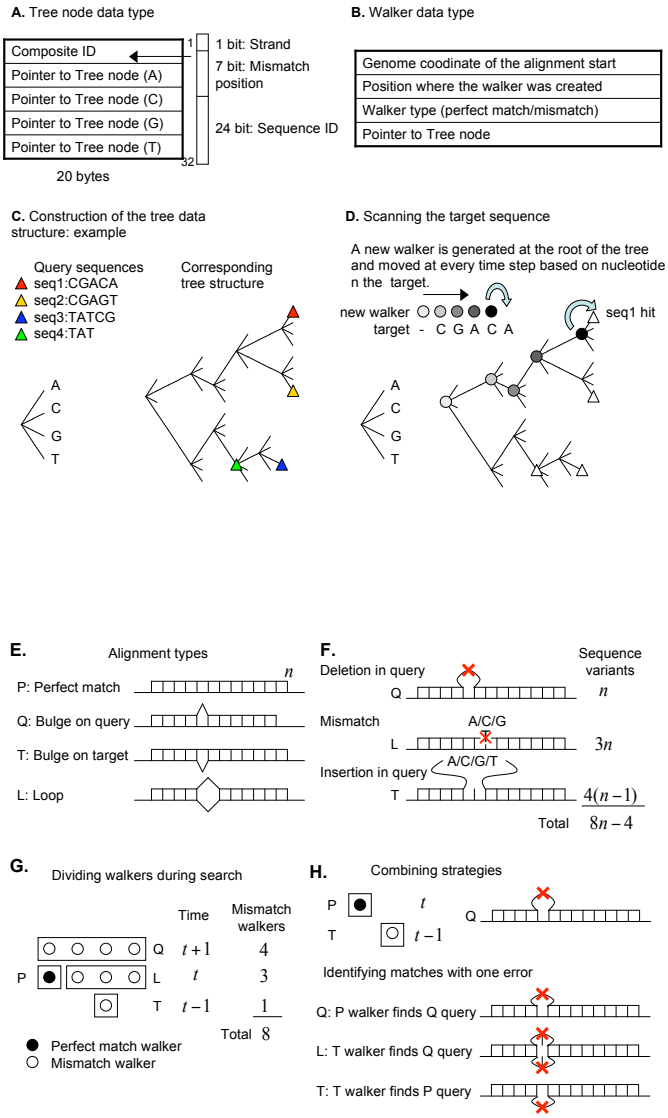


Figure 2.2: Schematic representation of the mapping algorithm.

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:

$$\begin{aligned} S &= L\left(1-\left(1-\frac{1}{A^L}\right)^N\right) \\ &+ \sum_{l=1}^{L-1} l\left[\left(1-\left(1-\frac{1}{A^l}\right)^N\right)-\left(1-\left(1-\frac{1}{A^{l+1}}^N\right)\right)\right] \\ &= L\left(1-\left(1-\frac{1}{A^L}\right)^N\right)+\sum_{l=1}^{L-1} l\left(\left(1-\frac{1}{A^{l+1}}^N\right)-\left(1-\frac{1}{A^l}\right)^N\right) \\ &= L-\sum_{l=1}^{L}(1-\frac{1}{A^l})^N. \end{aligned} \tag{2.1}$$

As shown in Figure 2.3, 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 2.2F). 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 2.2G). 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 2.2H): 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

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

The behavior of these functions of N are shown in Figure 2.3A for $A=4$ and $L=16,20,24,28,32,36,40$.

2.2.4 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 [176]. 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 2.3C 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.shtml.

2.3 Automated annotation of 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 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

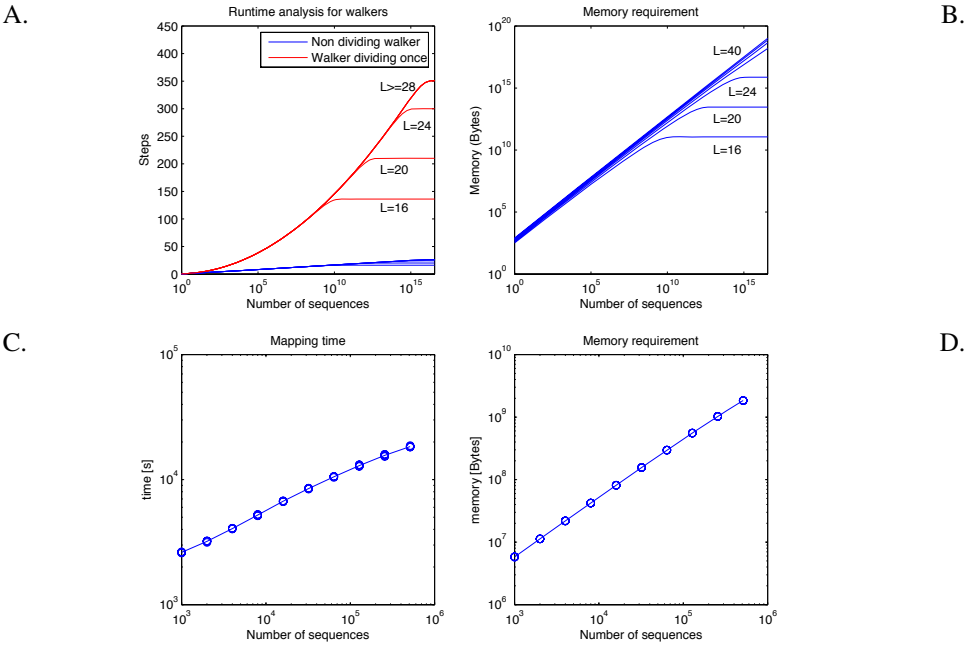


Figure 2.3: Performance of the mapping algorithm. A: estimated average number of steps performed by a walker as a function of the number of small RNAs represented in the tree. Blue corresponds to the case of perfect matches only, red to perfect and 1-error matches. The alphabet size was $A = 4$. The 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.

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 [177]. For this reason we use both the genome annotation as well as mappings of the small RNAs to transcripts with known function to functionally annotate small RNAs. We download 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 use 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"
- piRNA - Genbank sequence search and data from Aravin et al. [113]
- mRNA - Genbank sequence search using human/mouse/rat as species and mRNA as "Molecule type" to filter the records.

For our recent analysis of mammalian miRNA expression [170] we additionally curated the set of human, mouse and rat miRNAs that have been described to date, taking special care to identify the orthologs of all known miRNAs in these three species.

After compiling these data, we can proceed to annotate small RNAs in individual samples. The individual steps of the computational annotation and the software tools (other than custom Perl scripts) that we use in each of them are as follows:

1. Map small RNAs to genome using oligomap ($0-/1-$error matches) and WU-BLAST (matches with ≥ 2 errors).
2. For each small RNA identify the locus/loci with minimum number of errors (mismatch, insertion, deletions) in the small RNA-to-genome mapping.
3. Filter out too distant mappings ($< 92.5\%$ identity).
4. Map small RNAs to annotated sequences using oligomap ($0-/1-$error matches), WU-BLAST (matches with ≥ 2 errors).
5. For each small RNA identify the sequences with minimum number of errors (mismatch, insertion, deletions) in the small RNA-to-annotated sequence mapping.

6. Filter out too distant mappings $< 92.5\%$ identity.

7. Assign a functional category to each small RNA based on all its best mappings.

8. Compute the number of sequences derived from each arm of each pre-miRNA.

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.

The incompleteness of databases and ambiguous mappings pose problems also in the annotation of miRNAs. For instance, until recently, the miRNA repository miRBase (http://microrna.sanger.ac.uk/sequences/) only contained miRNAs from individual publications. These studies did not generally go as far as identifying the homologs of the miRNAs that they uncovered in all sequenced genomes, and subsequent cloning studies frequently isolated such homologs. To be able to distinguish entirely novel miRNAs from homologs of miRNAs that were cloned previously from other species, we use in our annotation procedure all pre-miRNAs from miRBase, irrespective of the species. When no match to in-species precursors is found, or when a more precise match to an out-of-species precursor is found, the out-of-species precursor is used in the annotation of the small RNA.

In order to compare miRNA expression across samples, we need to determine the number of clones of each individual miRNA that have been sequenced. In some cases, a small RNA matches equally well multiple miRNA precursors, and we do not know from which of these precursors the small RNA originated. Assuming that any of these precursors was equally likely to give rise to the small RNA, we count each small RNA towards all the equally-well matched precursors, with a weight that is the inverse of the number of such precursors. We thus obtain sequence counts reflecting the expression of mature miRNAs and of miRNA precursors, which we can compare across samples.

We have combined all of these concepts into a software system for the annotation of small RNAs, which we used to construct a mammalian miRNA expression atlas [170]. Most recently, we have implemented this system as a web server (http://www.mirz.unibas.ch/smiRNA-annotation/) that allows a user to annotate small RNAs from a set of samples. The server is coupled to a database of publicly available miRNA expression profiles (currently the miRNA atlas [170]), which we will continue to update as more datasets become available. This enables users to analyze their own samples, starting with the extraction of small RNAs from sequence reads, and continuing with the visualization and comparison of miRNA expression in these samples relative to all others that are publicly available.

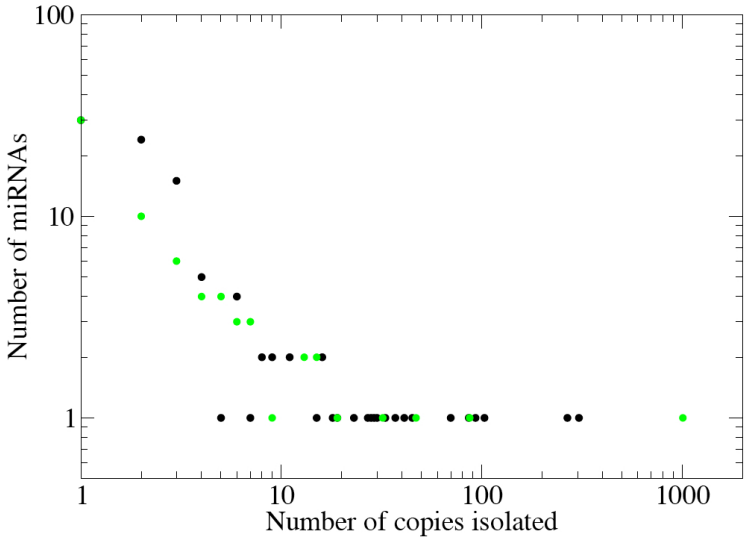


Figure 2.4: Frequency distribution of miRNA counts in 2 individual samples from frontal cortex (black) and liver (green). Most miRNAs in a given sample occur at low frequency, with a few miRNAs having relatively high frequency.

2.4 Comparison of miRNA expression profiles

2.4.1 Clustering samples

One of the main goals of small RNA cloning is to characterize miRNA expression across tissues and to understand the changes that take place during cell differentiation or during pathogenic processes. One approach to these questions is to identify significant changes in miRNA expression between samples. Because in general we do not have absolute measurements of miRNA expression, but only the relative counts of different miRNAs within a sample, what we can detect and quantify are changes in the *relative frequencies* of miRNAs within samples. Compared to other technologies used to measure gene expression, such as microarrays, our data consisted up to now of relatively small RNA libraries, containing on the order of a thousand clones. Interestingly, we typically find frequency distributions in which a few miRNAs are highly expressed, occurring in hundreds of copies, while most miRNAs occur in only a handful of copies (Figure 2.4). Given the relatively small sample sizes, the miRNA copy numbers will be subject to large sampling noise. To identify significant changes in miRNA expression we adopted a Bayesian probability framework.

We want to quantify the overall similarity of the miRNA expression profiles of two samples. In each sample there is a true, but unknown, distribution of frequencies p_i for each of the miRNAs i. Let p_i denote the true frequency of miRNA i in the first sample, and let q_i denote the true frequency of miRNA i in the second sample.

The observed copy numbers n_i and m_i of each miRNA i in the two samples can be considered multinomial samples from the distributions $\{p_i\}$ and $\{q_i\}$, and for a given set of frequencies the probability of the data is given by

$$P(\{n_i\},\{m_i\}|\{p_i\},\{q_i\}) = \prod_i [p_i^{n_i} q_i^{m_i}]\,. \tag{2.5}$$

We calculate the probability of the observed counts $\{n_i\}$ and $\{m_i\}$ under two models. The first model (model "I") assumes that the frequencies p_i and q_i are both unknown and independent of each other. To calculate the likelihood L_I of this model we assign a Dirichlet prior probability to the unknown frequency distributions, i.e. a prior of the form

$$P(\{p_i\}) = \Gamma(k\alpha)\prod_i \frac{p_i^{\alpha-1}}{\Gamma(\alpha)}, \tag{2.6}$$

where k is the number of miRNAs and α is the pseudo-count of the Dirichlet prior, and integrate over all possible distributions $\{p_i\}$ and $\{q_i\}$, i.e.

$$L_I = \int P(\{n_i\},\{m_i\}|\{p_i\},\{q_i\})P(\{p_i\})P(\{q_i\})dpdq, \tag{2.7}$$

where the integral is over all distributions $\sum_i p_i = \sum_i q_i = 1$. The integral can be performed analytically and we obtain

$$L_I = \frac{\Gamma(k\alpha)^2}{\Gamma(n+k\alpha)\Gamma(m+k\alpha)}\prod_i \frac{\Gamma(n_i+\alpha)\Gamma(m_i+\alpha)}{\Gamma(\alpha)^2}, \tag{2.8}$$

where n is the total number of miRNAs in the first sample and m the total number of miRNAs in the second sample.

The second model assumes that the relative frequency of any miRNA i is the same between the two samples ("S" model), i.e. it assumes $p_i = q_i$ for all i. The likelihood L_S of this model is given by

$$\begin{aligned} L_S &= \int P(\{n_i\},\{m_i\}|\{p_i\},\{p_i\})P(\{p_i\})dp \\ &= \frac{\Gamma(k\alpha)}{\Gamma(n+m+k\alpha)}\prod_i \frac{\Gamma(n_i+m_i+\alpha)}{\Gamma(\alpha)}. \end{aligned} \tag{2.9}$$

Finally, the posterior probability of the S model is given by $L_S/(L_I+L_S)$ and we can use this to define a *distance* $d = \log(\frac{L_I+L_S}{L_S})$ between the expression profiles of the two samples, which we can further use for the hierarchical clustering of samples of miRNAs. Note that as the posterior probability of the S model goes to 1 the distance goes to zero, and that as the posterior probability of the S model goes to zero, the distance goes to infinity.

2.4.2 Clustering miRNAs

Instead of clustering samples based on the overall expression profile of the miRNAs we can also cluster groups of miRNAs into clusters based on the expression profiles of the miRNAs across samples. We consider all ways in which miRNAs can be partitioned into clusters. For each such a partition we can calculate a likelihood as follows. Let c denote a cluster in the partition, $|c|$ the number of miRNAs in this partition, n_c^i the total count of miRNAs from cluster c in sample i, and ρ_c^i the (unknown) overall frequency of cluster c in sample i. We will again use a Dirichlet prior over the unknown cluster frequencies ρ_c^i and calculate the likelihood of the partition by integrating over all possible distribution ρ_c^i (separately for each sample i). In addition, we will assume that each miRNA in cluster c will have the *same* frequency $\rho_c^i/|c|$ in sample i. That is, we treat the relative frequencies of the different clusters in each of the samples as unknown variables but demand that within each cluster all miRNAs have the same frequency. Under this model we obtain for the likelihood

$$\begin{aligned} L &= \prod_i \left[\int \frac{\Gamma(k\alpha)}{\Gamma(\alpha)^k} \prod_c \left(\frac{\rho_c^i}{|c|} \right)^{\alpha-1+n_c^i} d\rho^i \right] \\ &= \prod_i \left[\frac{\Gamma(k\alpha)}{\Gamma(n^i+k\alpha)} \prod_c \left(\frac{\Gamma(n_c^i+\alpha)}{\Gamma(\alpha)} |c|^{1-\alpha-n_c^i} \right) \right], \end{aligned} \tag{2.10}$$

where the first product is over all samples, the integral for each sample is over the unknown frequencies ρ_c^i for that sample (with $\sum_c \rho_c^i = 1$), k is the number of clusters in the partition k, the second product is over all k clusters in the partition, and the sum in the exponent is over all miRNAs a that belong to cluster c.

We then cluster miRNAs hierarchically: we start with each miRNA being placed in its own cluster, and then at every step, we consider merging two clusters together. To decide which clusters to merge, we use the same concept as above: we denote the partition in which the two clusters are merged as the "S" model, and the partition in which the two clusters are separate as the "I" model. We then cluster at each step the two clusters for which the likelihood ratio L_S/L_I is maximal. Consider two clusters c and c'. Using expression (2.10) we have for the likelihood ratio

$$\frac{L_S}{L_I} = \prod_i \left[\frac{|c|^{n_c^i+\alpha-1}|c'|^{n_{c'}^i+\alpha-1}}{(|c|+|c'|)^{n_c^i+n_{c'}^i+\alpha-1}} \frac{\Gamma(n_c^i+n_{c'}^i+\alpha)\Gamma(\alpha)}{\Gamma(n_c^i+\alpha)\Gamma(n_{c'}^i+\alpha)} \frac{\Gamma(n^i+k\alpha)\Gamma((k-1)\alpha)}{\Gamma(n^i+(k-1)\alpha)\Gamma(k\alpha)} \right]. \tag{2.11}$$

These clustering algorithms have been used in our analysis of the miRNA atlas data [170].

2.5 Concluding remarks

Many researchers are using small RNA cloning and sequencing to study gene expression of both small RNAs and mRNAs. The first step in the analysis of such data sets is the mapping of small RNAs to both genome and to sequences with known function. We developed an algorithm that allows very rapid mapping of these small RNAs assuming that only very close matches (with 0 or 1 error) are desired. We further developed a web server that applies the analysis steps that we have used in constructing the miRNA atlas to provide a functional annotation for user-provided data sets. One of the most typically asked questions is how small RNAs samples differ, and which of the small RNAs are most responsible for the difference. Here we described the Bayesian framework that we currently use in our server to identify miRNAs whose relative frequency between changes most significantly between samples, taking into account the noise inherent in the relatively small counts of these molecules in typical samples.

Chapter 3

MirZ: An integrated microRNA expression atlas and target prediction resource

Jean Hausser[1], Philipp Berninger[1], Christoph Rodak[1,2], Yvonne Jantscher[3], Stefan Wirth[4] and Mihaela Zavolan1[1,5]

1 Biozentrum, Universität Basel and Swiss Institute of Bioinformatics, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

2 Fachhochschule Nordwestschweiz, Schulthess-Allee 1, 5201 Brugg, Switzerland

3 Fachhochschule Hagenberg, Softwarepark 11, 4232 Hagenberg, Austria

4 Institut für Informatik, Universität Leipzig, Härtelstr. 16-18 04107 Leipzig, Germany

5 To whom correspondence should be addressed.

published in *Nucleic Acids Research 37:W266-W272, 2009*

MicroRNAs (miRNAs) are short RNAs that act as guides for the degradation and translational repression of protein-coding mRNAs. A large body of work showed that miRNAs are involved in the regulation of a broad range of biological functions, from development to cardiac and immune system function, to metabolism, to cancer. For most of the over 500 miRNAs that are encoded in the human genome the functions still remain to be uncovered. Identifying miRNAs whose expression changes between cell types or between normal and pathological conditions is an important step towards characterizing their function as is the prediction of mRNAs that could be targeted by these miRNAs. To provide the community the possibility of exploring interactively miRNA expression patterns and the can-

didate targets of miRNAs in an integrated environment, we developed the MirZ web server, which is accessible at www.mirz.unibas.ch. The server provides experimental and computational biologists with statistical analysis and data mining tools operating on up-to-date databases of sequencing-based miRNA expression profiles and of predicted miRNA target sites in species ranging from *Caenorhabditis elegans* to *Homo sapiens*.

3.1 Introduction

MicroRNAs (miRNA) are a continuously growing class of small RNAs that act as guides in the translational silencing and degradation of target mRNAs [178]. Many miRNAs are conserved over large evolutionary distances such as between human and worm [5]. Fundamental biological processes such as development [3,71–73], metabolism [75–77], cardiac [179] and immune system function [180] have been shown to be regulated by miRNAs, and aberrant miRNA expression has been associated with cancers [78, 181].

There are various approaches to miRNA expression profiling, one of which is small RNA sequencing. Classical cloning and sequencing of size-separated small RNAs have been used to generate a large atlas of miRNA expression profiles [170], and this approach can be scaled up considerably through deep sequencing technologies [172]. Microarray-based expression profiling is also a popular approach, which has been used for instance to characterize the miRNA expression cancer samples [78]. In contrast to sequencing, microarray-based profiling does not allow identification of novel miRNAs.

Numerous approaches have also been proposed for miRNA target prediction. Because the 5' end of miRNAs (known as "seed") appears to be important for target recognition, a number of tools focus on the evolutionary conservation of miRNA seed-complementary regions in 3'UTRs [65, 68, 182, 183]. Other approaches emphasize the energy of hybridization between miRNA and target [184–186], the expected anti-correlation between the expression level of miRNAs and their mRNA targets [187,188], the properties of the environment of the miRNA target site [69, 189], or combine various features of the miRNA target site itself [190, 191].

Studies in both native expression [192] as well as transfection-induced miRNA overexpression situations [74, 81] indicate that within a given tissue, the miRNAs that are most strongly expressed have the largest impact on mRNA targets. For this reason, deciphering the miRNA-dependent post-transcriptional regulatory layer in a given tissue or cell type needs to start from the miRNA expression profile of that tissue or cell type. Conversely, it is very common that one identifies differences in miRNA expression between cells at various stages of differentiation or between normal and malignant cells, and the natural question is what mRNAs are most likely to be affected by the change in miRNA expression. To address these types of questions, we developed

MirZ (www.mirz.unibas.ch), a web service that integrates two resources that we developed in the context of previous research projects: the smiRNAdb miRNA expression atlas [170], and the ElMMo miRNA target prediction algorithm [68].

3.2 Materials and methods

3.2.1 The smiRNAdb miRNA expression atlas

smiRNAdb [170] is a unique, web-accessible and widely used resource of miRNA profiles determined by sequencing from hundreds of *Homo sapiens*, *Mus musculus* and *Rattus norvegicus* samples. The web interface of smiRNAdb features an extended repertoire of on-line analyses such as visualization and hierarchical clustering of miRNA expression profiles, principal component analysis, comparison of miRNA expression between two (sets of) samples with the aim of identifying the miRNAs whose expression differs most between the samples. We used the Brenda tissue ontology [193, 194] as a guide in organizing the samples such that the user can readily identify related cell lineages or normal and pathological samples derived from a given tissue type. Our tissue hierarchy has four levels: the organ/system (e.g. hematopoietic system), subsystem (e.g. lymphoid lineage), cell type (e.g. B cell), further cell type classification (e.g. B lymphocyte). MiRNAs themselves can be analyzed independently, grouped by their 2-7 subsequence, or grouped in precursor clusters. Two miRNAs are placed in the same precursor cluster if their loci are within 50 kilobases of each other in the genome, or if they share a mature form.

As an example, one may be interested in comparing miRNA expression between effector and naive human $CD4^+$ T-lymphocytes. SmiRNAdb features a "Sample comparison" tool which was specifically designed for the pairwise comparison of miRNA (sets of) samples. The user would select to compare the sample named "hsa_T-cell-CD4-effector" to the sample named "hsa_T-cell-CD4-naive". Because the naive $CD4^+$ T cell sample and the effector $CD4^+$ T cell sample differ widely in the total number of sequenced miRNAs (1374 vs 89), the precision of the miRNA frequency estimates in the two samples will also be very different. This situation is common in sequencing-based datasets making the identification of miRNAs whose expression is *significantly* different a non-trivial problem. At the heart of the tools offered by smiRNAdb however, is a Bayesian model for computing the posterior probability that the *frequency* of a miRNA in the total miRNA population differs between two (sets of) samples. We compute this probability assuming a binomial sampling model and integrating over the unknown miRNA frequencies in the samples. This approach — described in details in Berninger et al. [195] — takes into account both the variability between sample sizes and the absolute miRNA counts.

Figure 3.1 shows the results of comparing the miRNA expression profiles of naive

vs effector CD4$^+$ cells. The names and sizes of the samples being compared are shown at the top of the page, followed by the log-likelihood ratio $\log(P_{\text{same}}/P_{\text{diff}})$ of two models, one that assumes that the frequencies of miRNAs are the same and one that assumes that they can be different between the samples. The log-likelihood ratio takes positive values when the miRNA frequencies are similar and negative values when they are different. In this case, the log-likelihood ratio is positive, indicating that overall, the frequencies of miRNAs in these samples are more likely to have been the same. The list of miRNAs ranked from most dissimilar to most similar expression follows. Each row contains the name of a miRNA, the direction of regulation (up or down), the cloning counts and frequencies in both samples, and provides a direct link to the predicted targets of the miRNA. The model indicates that with a 18% vs 54% cloning frequency, and despite the small size of the effector CD4$^+$ T cell sample, miR-142-5p is very likely to be down-regulated in effector cells. Again, this can be inferred from the negative value of $\log(P_{\text{same}}/P_{\text{diff}})$ for miR-142-5p. From this page, the user can select one or several miRNAs that came out differentially expressed and can browse the list of predicted targets (figure 3.2). In the case of miR-142-5p, the top 10 predicted targets include four transcription factors (AFF4, ONECUT2, ZFPM2 and ZNF148), and a kinase (PRPF4B) involved in pre-RNA splicing. These genes could provide a starting point for experimental studies on the function of miR-142-5p in T lymphocytes.

Since the original release of smiRNAdb, we have implemented an additional tool for performing principal component analysis on the miRNA expression profiles, we added more possibilities for the user to download miRNA profile data for further processing, and we started to incorporate other publicly available small RNA sequencing data sets from *Danio rerio*, *Drosophila melanogaster* and *Caenorhabditis elegans*. We reimplemented the software that was originally written in Perl CGI to use Java Server Faces technology and Apache / Tomcat. The computations are now performed on a computing cluster, with job distribution managed by the Sun Grid Engine queuing system. Finally, we enhanced the result screens of our on-line analysis tools with hyperlinks which directly take the user to the miRNA target predictions within the context of the smiRNAdb query, *i.e.* preserving the selected organism, miRNAs, and tissue (if available). Please refer to the web connectivity map in the supplementary material for an overview of the new links between smiRNAdb and ElMMo, as well as of the external ressources that we use in performing various analyses.

3.2.2 The ElMMo miRNA target prediction algorithm based on comparative genomic analysis

To be able to address the question of what mRNA is most likely affected by the change in expression of a miRNA, we coupled smiRNAdb to a PHP-based web interface to the

ElMMo miRNA target predictions [68].

Returning to the example of the hsa-miR-142-5p miRNA which was highlighted in section 3.2.1, the web interface allows aside from browsing the predicted targets, a number of other queries. For instance, given an organism (*Homo sapiens* in this example), the user can choose to scan for predicted miRNA target sites not only the default set of transcripts, which is all known RefSeq [196] mRNAs in the chosen organism, but also subsets of transcripts. The SymAtlas project [197] of the Genomics Institute of the Novartis Research Foundation (GNF) generated microarray-based *mRNA* expression profiles for a wide range of tissues. These profiles are incorporated in MirZ, giving the user the possibility to restrict target prediction to mRNAs that are expressed in a given cell type. The web interface further allows to scan an arbitrary number of mRNAs for up to 20 miRNAs simultaneously. Alternatively, the user can limit the number of mRNAs to scan to 20 mRNAs and then retrieve predicted target sites in these mRNAs for an arbitrary number of miRNAs.

MiRNAs exert their effector function through ribonucleoprotein complexes (miRNP) that contain, aside from the guiding miRNA a member of the Argonaute family of proteins. The determinants of productive miRNA-target site interactions are not entirely known, but a large body of work [30, 182, 198–201] established that perfect complementarity of the 7–8 nucleotides from the 5' end of the miRNA — the so-called miRNA "seed" — is critical for target recognition. Although miRNA target sites that do not satisfy this constraint have been described, at the genome-wide level the accuracy of predicting such sites is low [68, 182]. Other than perfect seed complementarity, the location of the putative target site within the 3' UTR [68, 69, 202], structural accessibility [185, 186, 203], the nucleotide composition in its vicinity [69, 189] and the complementarity of specific positions in the miRNA 3' end to the target site [69] have all been reported to improve the accuracy of miRNA target prediction, yet the relative importance of these features remains unknown. The ElMMo miRNA target prediction method that we developed is based on a Bayesian model that only uses comparative genomics information. Yet it has as high an accuracy as other widely used target prediction programs that incorporate additional constraints, and measures of predictive performance on a set of experimentally validated miRNA targets in *Drosophila melanogaster* can be found in the article describing the ElMMo method [68]. Importantly, our model does not have any free parameters, and can easily accommodate additional species whose genome sequence becomes available.

Going back to our example, figure 3.2 shows the ElMMo predictions for miR-142-5p in *Homo sapiens*. This result screen is organized in two sections: (1) a miRNA-centric summary featuring per-miRNA target prediction statistics and a figure showing the smiRNAdb tissues where the selected miRNAs are mostly expressed, and (2) a mRNA-centric summary that ranks all mRNAs predicted to be targeted by the selected miRNAs. In this later section, mRNAs are ordered by decreasing expected number

of miRNA target sites under selective pressure, defined as the sum of all target site posterior probabilities for the selected miRNAs. The location of the putative target sites in the 3'UTR is also indicated.

From the result screen, the user has the possibility to zoom onto a specific transcript to visualize the multiple genome alignments in the regions of the predicted target sites, and to find additional information about the targeted mRNAs from the Genbank database of the National Center for Biomedical Information (NCBI). Our web service also offers the possibility to run a Gene Ontology (GO) analysis searching for GO terms that are significantly over- or under-represented in the predicted miRNA targets through a modified version of the GeneMerge software [204]. For instance, in the case of miR-142-5p, the most significantly enriched Biological Process GO term is "regulation of transcription, DNA-dependent" (hypergeometric p-value $< 10^{-10}$, after Bonferroni multiple testing correction), followed by two "muscarinic acetylcholine receptor"-associated GO terms ($p < 10^{-10}$). The muscarinic acetylcholine receptor has been shown to be involved in autocrine control of cell proliferation, including the proliferation of immune cells [205]. This type of analyses could thus provide experimental scientists with clues to the function of miR-142-5p in the naive $CD4^+$ T cells.

The current release of ElMMo features miRNA target predictions for *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Danio rerio*, *Caenorhabditis elegans* and *Drosophila melanogaster*. Of these, the *Mus musculus* and *Rattus norvegicus* predictions were not present in our initial publication [68]. Furthermore, for the remaining organisms, the current predictions are based on the genome sequences of a larger set of species, because more fully-sequenced genomes became available since 2007. We further based our predictions on the most recent mRNA sequences and 3'UTR annotations provided by the RefSeq database [196]. Concerning the microarray profiles that the user can use to guide miRNA target discovery in specific tissues and aside from the *Homo sapiens* profiles that were used in our original ElMMo release [68], we incorporated similar mRNA expression profiles for *Mus musculus* and *Rattus norvegicus*. Finally, the ElMMo web interface now informs the user about the smiRNAdb samples in which the selected miRNAs are most strongly expressed.

3.2.3 Experimental data

The miRNA sequences that were used for miRNA sample annotation and for miRNA target prediction were obtained from the miRBase release 12.0 [206]. For the miRNA profiles, MirZ includes a total of 297 samples: 173 for *Homo sapiens* [170], 88 for *Mus musculus* [170], 16 for *Rattus norvegicus* [170], 10 for *Drosophila melanogaster* [123], 9 for *Danio rerio* [207], and 1 for *Caenorhabditis elegans* [96].

For miRNA target predictions, we used the most recent genome assemblies available at the University of California Santa Cruz (UCSC) [208]: hg18 for *Homo sapiens*,

mm9 for *Mus musculus*, rn4 for *Rattus norvegicus*, danRer5 for *Danio rerio*, ce6 for *Caenorhabditis elegans* and dm3 for *Drosophila melanogaster*. We further used the following UCSC genome assemblies in the pairwise genome alignments: panTro2, rheMac2, mm9, rn4, canFam2, monDom4, bosTau4 and galGal3 for *Homo sapiens*; panTro2, rheMac2, hg18, rn4, canFam2, monDom4, bosTau4 and galGal3 for *Mus musculus*; panTro2, rheMac2, hg18, mm9, canFam2, monDom4, bosTau3 and galGal3 for *Rattus norvegicus*; tetNig1, fr2 and oryLat2 for *Danio rerio*; caeJap1, caePb2, caeRem3, cb3 and priPac1 for *Caenorhabditis elegans*; dp4, droAna3, droEre2, droGri2, droMoj3, droPer1, droSec1, droSim1, droVir3, droWil1 and droYak2 for *Drosophila melanogaster*. mRNAs for all organisms were downloaded from the RefSeq database on January 21st 2009.

The links between sequence entities in various databases was made by mapping them all to the Gene database of NCBI [196]. MiRNA expression profiles, microarray mRNA profiles and miRNA target predictions are stored as relational databases managed by a PostgreSQL server (www.postgresql.org).

3.3 Conclusion and future directions

Using a concrete example comparing effector to naive $CD4^+$ T-cells, we showed how MirZ can help isolating miRNAs that may be involved in a given biological function, and then provide clues into which molecular pathways may be controlled by these miRNAs to achieve their biological function. The integration of miRNA expression profiles with genome-wide miRNA target prediction combined with the tools we implemented — a Bayesian model for sample comparison, multivariate exploratory statistics, GO-term enrichment analysis — makes MirZ a powerful tool for studying miRNA-based regulation.

Since its publication, the miRNA expression atlas has been a valuable resource to the research community, and with the more general availability of deep sequencing technologies, more miRNA expression data sets are expected to emerge. Being able to explore and compare these data sets in a unified framework is highly desirable, and we plan to further support such analyses by updating MirZ as new data sets become available. Particularly for *Drosophila melanogaster*, we currently only incorporate small-sized samples, and for *Caenorhabditis elegans* a whole-worm sample.

The target prediction methods also continue to evolve. In particular, additional determinants of miRNA targeting specificity must exist because not all transcripts that contain miRNA seed matches respond in a given experiment, but what these determinants are is still an open question [69, 189]. If a significantly better target prediction method emerges, this could be incorporated in our server.

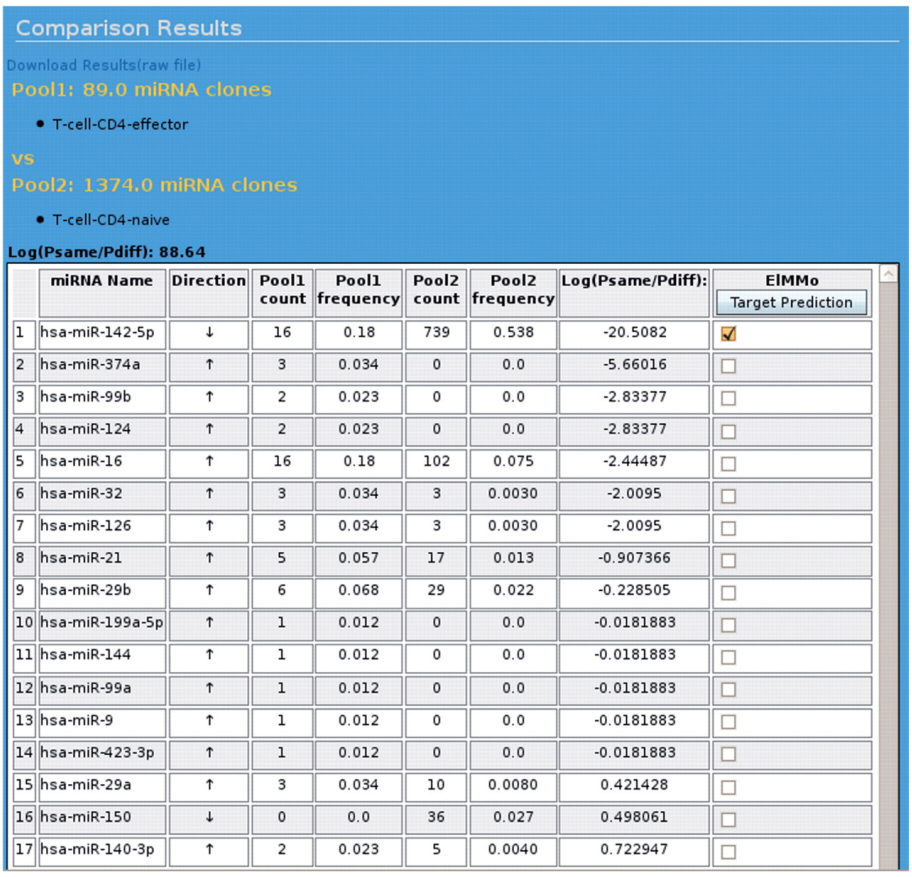
3.4 Funding

This work was supported by the Swiss National Fund grant #3100A0-114001 to M.Z. and by the Swiss Institute of Bioinformatics.

3.5 Acknowledgements

We acknowledge the contribution of Robin Vobruba and Philip Handschin from Fachhochschule Nordwestschweiz (Brugg, Switzerland) to the re-implementation of smiRNAdb in Java.

We also acknowledge the International Chicken Genome Sequencing Consortium [209] for the *Gallus gallus* genome, the Chimpanzee Genome Sequencing Consortium for the *Pan troglodytes* genome, the Baylor College of Medicine Human Genome Sequencing Center and the Rhesus Macaque Genome Sequencing Consortium for the *Macaca mulatta* genome, the Mouse Genome Sequencing Consortium for the *Mus musculus* genome [210], the Rat Genome Project at the Baylor College of Medicine Human Genome Sequencing Center for the *Rattus norvegicus* genome [211, 212], the Dog Genome Sequencing Project for the *Canis familiaris* genome [213], the Broad Institute for the *Monodelphis domestica* (opossum) genome, the Baylor College of Medicine Human Genome Sequencing Center for the *Bos taurus* and *Drosophila pseudoobscura* genomes [214], the Genoscope for the *Tetraodon nigroviridis* genome, the Morishita laboratory for the *Oryzias latipes* (Medaka) genome [215], WormBase for the *Caenorhabditis elegans* genome [216], the Agencourt Bioscience Corporation for the *Drosophila ananassae* and *Drosophila erecta* genomes, and the Drosophila 12 Genomes Consortium for all other *Drosophila* genomes [217]. The *Takifugu rubripes* genome was provided freely by the Fugu Genome Consortium for use in this publication only. The *C. brenneri, C. briggsae, C. japonica, C. remanei* and *P. pacificus* genomes were produced by the Genome Sequencing Center at Washington University School of Medicine in St. Louis and can be obtained from ftp://genome.wustl.edu/pub/organism/Invertebrates/.



Comparison Results

Download Results(raw file)

Pool1: 89.0 miRNA clones

- T-cell-CD4-effector

vs

Pool2: 1374.0 miRNA clones

- T-cell-CD4-naive

Log(Psame/Pdiff): 88.64

	miRNA Name	Direction	Pool1 count	Pool1 frequency	Pool2 count	Pool2 frequency	Log(Psame/Pdiff):	ElMMo Target Prediction
1	hsa-miR-142-5p	↓	16	0.18	739	0.538	-20.5082	☑
2	hsa-miR-374a	↑	3	0.034	0	0.0	-5.66016	☐
3	hsa-miR-99b	↑	2	0.023	0	0.0	-2.83377	☐
4	hsa-miR-124	↑	2	0.023	0	0.0	-2.83377	☐
5	hsa-miR-16	↑	16	0.18	102	0.075	-2.44487	☐
6	hsa-miR-32	↑	3	0.034	3	0.0030	-2.0095	☐
7	hsa-miR-126	↑	3	0.034	3	0.0030	-2.0095	☐
8	hsa-miR-21	↑	5	0.057	17	0.013	-0.907366	☐
9	hsa-miR-29b	↑	6	0.068	29	0.022	-0.228505	☐
10	hsa-miR-199a-5p	↑	1	0.012	0	0.0	-0.0181883	☐
11	hsa-miR-144	↑	1	0.012	0	0.0	-0.0181883	☐
12	hsa-miR-99a	↑	1	0.012	0	0.0	-0.0181883	☐
13	hsa-miR-9	↑	1	0.012	0	0.0	-0.0181883	☐
14	hsa-miR-423-3p	↑	1	0.012	0	0.0	-0.0181883	☐
15	hsa-miR-29a	↑	3	0.034	10	0.0080	0.421428	☐
16	hsa-miR-150	↓	0	0.0	36	0.027	0.498061	☐
17	hsa-miR-140-3p	↑	2	0.023	5	0.0040	0.722947	☐

Figure 3.1: Screenshot of the web page showing the result from comparing miRNA expression of human $CD4^+$ effector T cells with the $CD4^+$ naive T cells. Details are provided in the text

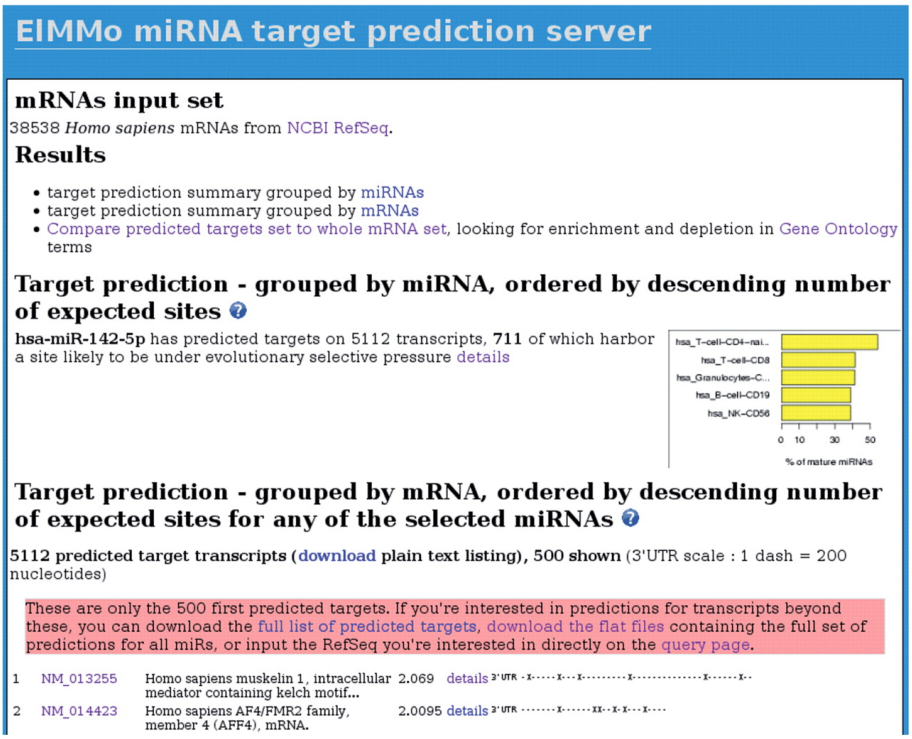


Figure 3.2: Screenshot of the web page showing the ElMMo miRNA target predictions for miR-142-5p in all *Homo sapiens* RefSeq mRNAs. The target predictions results are organized in two sections. The first section — located on the upper part of the web page — is miRNA-centric and features miRNA target predictions statistics as well as a figure showing the smiRNAdb tissues where the miRNA is mostly expressed. The second, mRNA-centric section is located on the lower part of the web page and provides a ranked list of mRNA predicted to be targeted by miR-142-5p.

Chapter 4

Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP

Markus Hafner[1,5], Markus Landthaler[1,4,5], Lukas Burger[2], Mohsen Khorshid[2], Jean Hausser[2], Philipp Berninger[2], Andrea Rothballer[1], Manuel Ascano, Jr.[1], Anna-Carina Jungkamp[1,4], Mathias Munschauer[1], Alexander Ulrich[1], Greg S. Wardle[1], Scott Dewell[3], Mihaela Zavolan[2,6] and Thomas Tuschl[1,6]

1 Howard Hughes Medical Institute, Laboratory of RNA Molecular Biology, The Rockefeller University, 1230 York Avenue, Box 186, New York, NY 10065, USA

2 Biozentrum der Universität Basel and Swiss Institute of Bioinformatics (SIB), Klingelbergstr. 50-70, CH-4056 Basel, Switzerland

3 Genomics Resource Center, The Rockefeller University, 1230 York Avenue, Box 241, New York, NY 10065, USA

4 Present address: Berlin Institute for Medical Systems Biology, Max-Delbrück-Center for Molecular Medicine, 13125 Berlin, Germany

5 These authors contributed equally to this work

6 Correspondence: mihaela.zavolan@unibas.ch (M.Z.), ttuschl@rockefeller.edu (T.T.)

published in *Cell 141(1), 129-141, 2010*

RNA transcripts are subject to posttranscriptional gene regulation involving hundreds of RNA-binding proteins (RBPs) and microRNA-containing ribonucleoprotein complexes (miRNPs) expressed in a cell-type dependent fashion. We developed a cellbased crosslinking approach to deter-

mine at high resolution and transcriptome-wide the binding sites of cellular RBPs and miRNPs. The crosslinked sites are revealed by thymidine to cytidine transitions in the cDNAs prepared from immunopurified RNPs of 4-thiouridine-treated cells. We determined the binding sites and regulatory consequences for several intensely studied RBPs and miRNPs, including PUM2, QKI, IGF2BP1-3, AGO/EIF2C1-4 and TNRC6A-C. Our study revealed that these factors bind thousands of sites containing defined sequence motifs and have distinct preferences for exonic versus intronic or coding versus untranslated transcript regions. The precise mapping of binding sites across the transcriptome will be critical to the interpretation of the rapidly emerging data on genetic variation between individuals and how these variations contribute to complex genetic diseases.

4.1 Introduction

Gene expression in eukaryotes is extensively controlled at the posttranscriptional level by RNA-binding proteins (RBPs) and ribonucleoprotein complexes (RNPs) modulating the maturation, stability, transport, editing and translation of RNA transcripts [218–220]. Vertebrate genomes encode several hundred RBPs [221], each containing one or more domains able to specifically recognize target transcripts. Furthermore, hundreds of microRNAs (miRNAs) bound by Argonaute (AGO/EIF2C) proteins mediate destabilization and/or inhibition of translation of partially complementary target mRNAs [67]. To understand how the interplay of these RNA-binding factors affects the regulation of individual transcripts, high resolution maps of in vivo protein-RNA interactions are necessary [222].

A combination of genetic, biochemical and computational approaches are typically applied to identify RNA-RBP or RNA-RNP interactions. Microarray profiling of RNAs associated with immunopurified RBPs (RIP-Chip) [223] defines targets at a transcriptome level, but its application is limited to the characterization of kinetically stable interactions and does not directly identify the RBP recognition element (RRE) within the long target RNA. Nevertheless, RREs with higher information content can be derived computationally from RIP-Chip data, e.g., for HuR [224] or for Pumilio [225].

More direct RBP target site information is obtained by combining in vivo UV crosslinking [226, 227] with immunoprecipitation [228, 229] followed by the isolation of crosslinked RNA segments and cDNA sequencing (CLIP) [230]. CLIP was used to identify targets of the splicing regulators NOVA1 [231], FOX2 [232] and SFRS1 [233] as well as U3 snoRNA and pre-rRNA [234], pri-miRNA targets for HNRNPA1 [235], EIF2C2/AGO2 protein binding sites [236] and ALG-1 target sites in *C. elegans* [237]. CLIP is limited by the low efficiency of UV 254 nm RNA-protein crosslinking, and the location of the crosslink is not readily identifiable within the sequenced crosslinked

fragments, raising the question of how to separate UV-crosslinked target RNA segments from background noncrosslinked RNA fragments also present in the sample.

Here, we describe an improved method for isolation of segments of RNA bound by RBPs or RNPs, referred to as PAR-CLIP (*P*hoto*ac*tivatable-*R*ibonucleoside-Enhanced *C*ross*l*inking and *I*mmuno*p*recipitation). To facilitate crosslinking, we incorporated 4-thiouridine (4SU) into transcripts of cultured cells and identified precisely the RBP binding sites by scoring for thymidine (T) to cytidine (C) transitions in the sequenced cDNA. We uncovered tens of thousands of binding sites for several important RBPs and RNPs and assessed the regulatory impact of binding on their targets. These findings underscore the complexity of posttranscriptional regulation of cellular systems.

4.2 Results

4.2.1 Photoactivatable Nucleosides Facilitate RNA-RBP Crosslinking in Cultured Cells

Random or site-specific incorporation of photoactivatable nucleoside analogs into RNA in vitro has been used to probe RBP- and RNP-RNA interactions [238,239]. Several of these photoactivatable nucleosides are readily taken up by cells without apparent toxicity and have been used for in vivo crosslinking [240]. We applied a subset of these nucleoside analogs (Figure 4.1A) to cultured cells expressing the FLAG/HA-tagged RBP IGF2BP1 followed by UV 365 nm irradiation. The crosslinked RNA-protein complexes were isolated by immunoprecipitation, and the covalently bound RNA was partially digested with RNase T1 and radiolabeled. Separation of the radiolabeled RNPs by SDS-PAGE indicated that 4SU-containing RNA crosslinked most efficiently to IGF2BP1. Compared to conventional UV 254 nm crosslinking, the photoactivatable nucleosides improved RNA recovery 100- to 1000-fold, using the same amount of radiation energy (Figure 4.1B). We refer to our method as PAR-CLIP (*P*hotoactiv*a*table-*R*ibonucleoside-Enhanced *C*ross*l*inking and *I*mmuno*p*recipitation) (Figure 4.1C).

We evaluated the cytotoxic effects upon exposure of HEK293 cells to 100 μM and 1 mM of 4SU or 6SG in tissue culture medium over a period of 12 hr by mRNA microarrays. The mRNA profiles of 4SU or 6SG treated cells were very similar to those of untreated cells, suggesting that the conditions for endogenous labeling of transcripts were not toxic.

To guide the development of bioinformatic methods for identification of binding sites, we first studied human Pumilio 2 (PUM2), a member of the Puf-protein family (Figure 4.2A) known for its highly sequence-specific RNA binding [241].

4.2.2 Identification of PUM2 mRNA Targets and Its RRE

PUM2 protein crosslinked well to 4SU-labeled cellular transcripts (Figure 4.2B). The crosslinked segments were converted into a cDNA library and Solexa sequenced [242]. The sequence reads were aligned against the human genome and EST databases. Reads mapping uniquely to the genome with up to one mismatch, insertion or deletion were used to build clusters of sequence reads (Figure 4.2C). We obtained 7523 clusters originating from about 3000 unique transcripts, 93% of which were found within the 3' untranslated region (UTR) in agreement with previous studies [243]. All sequence clusters with mapping and annotation information are available online (http://www.mirz.unibas.ch/restricted/clipdata/RESULTS/index.html).

PhyloGibbs analysis [244] of the top 100 most abundantly sequenced clusters, as expected, yielded the PUM2 RRE, UGUANAUA [245] (Figure 4.2D). Unexpectedly, over 70% of all sequence reads that gave rise to clusters showed a T to C mutation compared to the genome. Ranking of sequence read clusters according to the frequency of T to C mutation further enriched for the PUM2 RRE indicating that the T to C mutation is diagnostic of sequences interacting with the RBP. The T to C changes were not randomly distributed: the T corresponding to U7 of the RRE mutated at higher frequency compared to the Ts corresponding to U1 and U3 (Figure 4.2E). Our analyses suggest that the reverse transcriptase specifically misincorporated dG across from crosslinked 4SU residues and that local amino acid environment also affected crosslinking efficiency. Uridines proximal to the RRE also exhibited an increased T to C mutation frequency, indicating that crosslinks also form in close proximity to an RRE and that our method even captured PUM2 binding sites that did not have a U7 in its RRE.

4.2.3 Identification of QKI RNA Targets and Its RRE

To further validate our method, we applied it to the RBP Quaking (QKI), which contains a single heterogeneous nuclear ribonucleoprotein K homology (KH) domain (Figures 4.3A,B). The RRE ACUAAY was determined by SELEX [246], but in vivo targets are largely undefined. Mice with reduced expression of QKI show dysmyelination and develop rapid tremors or "quaking" 10 days after birth. Previous studies suggested that QKI participates in pre-mRNA splicing, mRNA export, mRNA stability and protein translation [247].

PhyloGibbs analysis of the 100 most abundantly sequenced clusters yielded the RRE AYUAAY (Figures 4.3C,D), similar to a motif identified by SELEX [246]. We found approximately 6000 clusters mapping to 2500 transcripts. Close to 75% of these clusters were derived from intronic sequences, supporting the hypothesis that QKI is a splicing regulator [247] and 70% of the remaining exonic clusters fall into 3' UTRs.

Mutation analysis of the clustered sequence reads showed that the T correspond-

ing to U2 in AUUAAY was frequently altered to C whereas the T corresponding to U3 in AUUAAY or ACUAAY remained unaltered (Figure 4.3E). Crosslinking of 4SU residues located in immediate vicinity to the RRE was mostly responsible for exposing the motif with C2, showing that crosslinking inside the recognition element is not a precondition for its identification. Hence, the discovery of RREs is unlikely to be prevented by sequence-dependent crosslinking biases as long as deep enough sequencing captures these interaction sites at and nearby the RRE.

4.2.4 T to C Mutations Occur at the Crosslinking Sites

To better characterize the T to C transition observed in crosslinked RNA segments, we UV 365 nm crosslinked oligoribonucleotides containing single 4SU substitutions to recombinant QKI (Figures 4.3F,G). The crosslinking efficiency varied 50-fold and mirrored the results of the mutational analysis (Figure 4.3G). The least effective crosslinking was observed for placement of 4SU at position 3 of the QKI RRE (4SU9), and the most effective crosslinking was found at position 2 of the QKI RRE (4SU10); the crosslinking efficiency for two positions outside of the RRE (4SU2 and 4SU4) was intermediate. Neither of these substitutions affected RNA-binding to recombinant QKI protein as determined by gel-shift analysis, whereas mutations of the recognition element weakened the binding between 2.5- and 9-fold.

Next, we sequenced libraries prepared from noncrosslinked as well as QKI-protein-crosslinked oligoribonucleotides containing 4SU at indicated positions (Figure 4.3F). The fraction of sequence reads with T to C changes obtained from nonirradiated 4SU-containing oligoribonucleotides varied between 10 and 20%, and increased to 50% to 80% upon crosslinking. The variation of the degree of T to C changes in the crosslinked samples is most likely determined by background of noncrosslinked oligoribonucleotides. Presumably, the T to C transition frequency is increased upon crosslinking as a direct consequence of a chemical structure change of the 4SU nucleobase upon crosslinking to protein amino acid side chains, resulting in altered stacking or hydrogen bond donor/acceptor properties directing the preferential incorporation of dG rather than dA during reverse transcription. At the doses of 4SU applied to cultured cells, about 1 out of 40 uridines was substituted by 4SU as determined by HPLC analysis of the nucleoside composition of total RNA. Assuming a 20% T to C conversion rate for a noncrosslinked 4SU-labeled site, we estimated that the average T to C conversion rate of 40-nt sequence reads derived from background noncrosslinked sequences will be near 5%. Clusters of sequence reads with average T to C conversion above this threshold, irrespective of the number of sequence reads, most certainly represent crosslinking sites. The ability to separate signal from noise by focusing on clusters with a high frequency of T to C mutations rather than clusters with the largest number of reads, represents a major enhancement of our method over UV 254 nm crosslinking

methods.

To assess whether the transcripts identified by PAR-CLIP are regulated by QKI, we analyzed the mRNA levels of mock-transfected and QKI-specific siRNA-transfected cells with microarrays. Transcripts crosslinked to QKI were significantly upregulated upon siRNA transfection, indicating that QKI negatively regulates bound mRNAs (Figure 4.3H), consistent with previous reports of QKI being a repressor [247].

4.2.5 Identification of IGF2BP Family RNA Targets and Its RRE

We then applied PAR-CLIP to the FLAG/HA-tagged insulin-like growth factor 2 mRNA-binding proteins 1, 2, and 3 (IGF2BP1-3) (Figures 4.4A,B), a family of highly conserved proteins that play a role in cell polarity and cell proliferation [248]. These proteins are predominantly expressed in the embryo and regulate mRNA stability, transport and translation. They are re-expressed in various cancers [249, 250] and IGF2BP2 has been associated with type-2 diabetes [251]. The IGF2BPs are highly similar and contain six canonical RNA-binding domains, two RNA recognition motifs (RRMs) and four KH domains (Figure 4.4A). Therefore, target recognition for this protein family appears complex, with only a small number of coding and noncoding RNA targets being known so far. A precise definition of the RREs is missing [248].

The three IGF2BPs recognized a highly similar set of target transcripts, suggesting similar and redundant functions. PhyloGibbs analysis of the clusters derived from mRNAs (Figure 4.4C) yielded the sequence CAUH (H = A, U, or C) as the only consensus recognition element (Figure 4.4D), contained in more than 75% of the top 1000 clusters for IGF2BP1, 2 or 3. In total, we identified over 100,000 sequence clusters recognized by the IGF2BP family that map to about 8,400 protein-coding transcripts. The annotation of the clusters was predominantly exonic (ca. 90%) with a slight preference for 3' UTR relative to coding sequence (CDS). The mutation frequency of all sequence tags containing the element CAUH (H = A, C, or U) showed that the crosslinked residue was positioned inside the motif, or in the immediate vicinity (Figure 4.4E). The consensus motif CAUH was found in more than 75% of the top 1000 targeted transcripts, followed in more than 30% by a second motif, predominantly within a distance of three to five nucleotides. In vitro binding assays showed that nucleotide changes of the CAUH motif decreased, but did not abolish the binding affinity (Figure 4.4F).

To test the influence of IGF2BPs on the stability of their interacting mRNAs, as reported previously for some targets [248], we simultaneously depleted all three IGF2BP family members using siRNAs and compared the cellular RNA from knockdown and mock-transfected cells on microarrays. The levels of transcripts identified by PAR-CLIP decreased in IGF2BP-depleted cells, indicating that IGF2BP proteins stabilize their target mRNAs. Moreover, transcripts that yielded clusters with the highest T to C mutation frequency were most destabilized (Figure 4.4G), indicating that the ranking

criterion that we derived based on the analysis of PUM2 and QKI data generalizes to other RBPs.

For comparison to conventional and high-throughput sequencing CLIP [230, 231], we also sequenced cDNA libraries prepared from UV 254 nm crosslinking. Of the 8,226 clusters identified by UV 254 nm crosslinking of IGF2BP1, 4,795 were found in the PAR-CLIP dataset. Although UV 254 nm crosslinking identified the identical segments of a target RNA as PAR-CLIP, the position of the crosslink could not be readily deduced, because no abundant diagnostic mutation was observed.

4.2.6 Identification of miRNA Targets by AGO and TNRC6 Family PAR-CLIP

To test our approach on RNP complexes, we selected the protein components mediating miRNA-guided target RNA recognition. In animal cells, miRNAs recognize their target mRNAs through base-pairing interactions involving mostly 6–8 nucleotides at the 5' end of the miRNA (the so called "seed") [67]. Target sites were thought to be predominantly located in the 3' UTRs of mRNAs, and computational miRNA target prediction methods frequently resort to identification of evolutionarily conserved sites that are located in 3' UTRs and are complementary to miRNA seed regions [67, 252].

We isolated mRNA fragments bound by miRNPs from HEK293 cell lines stably expressing FLAG/HA-tagged AGO or TNRC6 family proteins [192]. The AGO IPs revealed two prominent RNA-crosslinked bands of 100 and 200 kDa, representing AGO, and likely TNRC6 and/or DICER1 protein. The TNRC6 IPs showed one prominent RNA-crosslinked protein of 200 kDa (Figure 4.5A).

From clusters (Figure 4.5B) formed by at least 5 PAR-CLIP sequence reads and containing more than 20% T to C transitions, we extracted 41 nt long regions centered over the predominant T to C transition or crosslinking site. The length of the crosslink-centered regions (CCRs) was selected to include all possible registers of miRNA/target-RNA pairing interactions relative to the crosslinking site.

PAR-CLIP of individual AGO proteins yielded on average about 4000 clusters that overlapped, supporting our earlier observation that AGO1-4 bound similar sets of transcripts [192]. We therefore combined the sequence reads obtained from all AGO experiments, which yielded 17,319 clusters of sequence reads at a cut-off of 5 reads. These clusters distributed across 4647 transcripts with defined GeneIDs, corresponding to 21% of the 22,466 unique HEK293 transcripts that we identified by digital gene expression (DGE).

PAR-CLIP of individual TNRC6 proteins yielded on average about 600 clusters that also overlapped substantially, again consistent with our observation that TNRC6 family proteins bind similar transcripts [192]. We therefore combined all sequence reads from all TNRC6 experiments, yielding 1865 clusters and CCRs. More than 50%

of these TNRC6 CCRs fell within 25 nt of an AGO CCR, and 26% overlapped by at least 75%, indicating that AGO and TNRC6 members bind to the same sites.

4.2.7 Comparison of miRNA Profiles from AGO PAR-CLIP to Non-crosslinked miRNA profiles

To relate the potential miRNA-target-site–containing CCRs to the endogenously expressed miRNAs, we determined the miRNA profiles from total RNA isolated from HEK293 cells, and miRNAs isolated from noncrosslinked AGO1-4 IPs by Solexa sequencing [242], and compared them to the profile from the miRNAs present in the combined AGO1-4 PAR-CLIP library. miRNA profiles obtained from total RNA and IP of the four AGO proteins in noncrosslinked cells correlated well (Figure 4.5C) supporting our observation that AGO1-4 bind the same targets [192]. The most abundant among the 557 identified miRNAs and miRNAs* were miR-103 (7% of miRNA sequence reads), miR-93 (6.5%), and miR-19b (5.5%). The 25 and 100 most abundant miRNAs accounted for 72% and 95% of the total of miRNA sequence reads, respectively. Comparison of the miRNA profile derived from the combined AGO PAR-CLIP library with the combined noncrosslinked libraries showed a good correlation (Spearman correlation coefficient of 0.56, Figure 4.5C).

Importantly, in the AGO PAR-CLIP library, the majority of miRNA sequence reads derived from prototypical miRNAs [170] displayed T to C conversion near or above 50%. The T to C conversion was predominantly concentrated within positions 8 to 13 (Figure 4.5D), residing in the unpaired regions of the AGO protein ternary complex [29]. Five of the 100 most abundant miRNAs in HEK293 cells lack uridines at position 8–13, yet only 2 of those miRNAs, miR-374a and b, showed no crosslinking, because uridines at residues 14 and higher can still be crosslinked. This frequency of crosslinks was substantially lower in the miRNAs whose expression did not correlate between AGO-IP and AGO PAR-CLIP samples compared to the miRNAs whose expression correlated well.

4.2.8 mRNAs Interacting with AGOs Contain miRNA Seed Complementary Sequences

Independent of any pairing models for miRNAs and their targets, we first determined the enrichment of all 16,384 possible 7-mers within the 17,319 AGO CCRs, relative to random sequences with the same dinucleotide composition. The most significantly enriched 7-mers, except for a run of uridines, corresponded to the reverse complement of the seed region (position 2–8) of the most abundant HEK293 miRNAs, and they were most frequently positioned 1–2 nt downstream of the predominant crosslinking site within the CCRs (Figure 4.6A). This places the crosslinking site near the center of the

AGO-miRNA-target-RNA ternary complex, where the target RNA is proximal to the Piwi/RNase H domain of the AGO protein [29]. The polyuridine motif lies within the region of target RNA that may be able to basepair with the 3' half of miRNA loaded into AGO proteins [29, 253]. Therefore, these stretches of uridine may contribute directly to miRNA-target RNA hybridization or, as has been suggested previously, they may represent an independent determinant of miRNA targeting specificity [69, 254].

To further examine the positional dependence of target RNA crosslinking, we aligned the CCRs containing 7-mer seed complements to the 100 most abundant miRNAs and plotted the position-dependent frequency of finding a crosslinked position (Figure 4.6B). This identified two additional crosslinking regions, which correspond to the unpaired 5' and 3' ends of the target RNA exiting from the AGO ternary complex, indicating that the window size of 41 nt centered on the predominant crosslink position always included the miRNA-complementary sites.

We then computed the number of occurrences of miRNA-complementary sequences of various lengths in the CCRs and calculated their enrichment. The most significant enrichment was generally obtained with 8-mers that were complementary to miRNA seed regions (pos. 1–8). Inspection of the region between 3 nt upstream and 9 nt downstream of the predominant crosslinking site reveals that approximately 50% of the CCRs contain 6-mers corresponding to one of the top 100 expressed miRNAs, with a 1.5-fold enrichment over random 6-mers. Given that 6-mers still showed some degree of excess conservation in comparative genomics studies [68, 182] and that our analysis was focused on a narrow window directly downstream of the crosslinking site, our results suggest that the majority of the CCRs represent bona fide miRNA binding sites. Furthermore, the number of miRNA seed complements for all known miRNAs correlated well with the expression levels of miRNAs found in HEK293 cells, and less well with miRNA profiles of other tissue samples.

The nucleotide composition of CCRs that contained at least one 7-mer seed complementary to one of the top 100 expressed miRNA showed a slightly elevated U-content (approx. 30% U) compared to those CCRs not containing seed matches, which was expected from previous bioinformatic analyses of functional miRNA-binding sites.

4.2.9 Noncanonical and 3' End Pairing of miRNAs to their mRNA Targets Is Limited

Structural and biochemical studies of the ternary complex of *T. thermophilus* Ago, guide and target indicated that small bulges and mismatches could be accommodated in the seed pairing region within the target RNA strand ([29]). We therefore searched for putative target RNA binding sites that did not conform to the model of perfect miRNA seed pairing, but rather contained a discontinuous segment of sequence complementarity to either target or miRNA with a minimum of 6 base pairs. We only

considered pairing patterns if they were significantly enriched in CCRs compared to dinucleotide randomized sequences, and if the CCRs containing them did not at the same time contain perfectly pairing seed-type sites. We identified 891 CCRs with mismatches and 256 with bulges in the seed region. Mismatches occurred most frequently across from position 5 of the miRNA as G-U or U-G wobbles, U-U mismatches and A-G mismatches (A residing in the miRNA). Therefore, it appears that only a small fraction of the miRNA target sites that we isolated (less than 6.6%), contained bulges or loops in the seed region.

To assess the role of auxiliary base pairing outside of the seed region, we selected CCRs that contained a 7-mer seed match to one of the 100 most abundant miRNAs. Supporting earlier computational results [69], we also detected a weak signal for contiguous 4-nt long matches to positions 13–15 of the miRNA (Figure 4.6C).

4.2.10 miRNA Binding Sites in CDS and 3' UTR Destabilize Target mRNAs to Different Degrees

The majority (84%) of AGO CCRs originated in exonic regions, with only 14% from intronic, and 2% from undefined regions. Of the exonic CCRs, 4% corresponded to 5' UTRs, 50% to CDS, and 46% to 3' UTRs (Figure 4.6D).

Evidence of widespread binding of miRNAs to the CDS was reported before [182, 255]. However, miRNAs are believed to predominantly act on 3' UTRs [67], with relatively few reports providing experimental evidence for miRNA-binding to individual 5' UTRs or CDS [255–259].

To obtain evidence that AGO CCRs indeed contain functional miRNA-binding sites, we blocked 25 of the most abundant miRNAs in HEK293 cells (Figure 4.5C) by transfection of a cocktail of 2'-O-methyl-modified antisense oligoribonucleotides and monitored the changes in mRNA stability by microarrays (Figure 4.7A). Consistent with previous studies of individual miRNAs [69], the magnitude of the destabilization effects of transcripts containing at least one CCR depended on the length of the seed-complementary region and dropped from 9-mer to 8-mer to 7-mer to 6-mer matches (Figure 4.7B). We did not find evidence for significant destabilization of transcripts that only contained imperfectly paired seed regions.

Next, we examined whether the change in stability of CCR-containing transcripts correlated with the number of binding sites. We found that multiple sites were more destabilizing compared to single sites (Figure 4.7C), and that multiple binding sites may also reside within a single 41-nt CCR. Both of these findings are in agreement with previous observations [69].

Then we analyzed the impact on stability for transcripts with CCRs exclusively present either in the CDS or the 3' UTR; there were not enough transcripts to assess the impact of CCRs derived from the 5' UTR. CDS-localized sites only marginally

reduced mRNA stability (Figure 4.7D), independent of the extent of seed pairing. To gain more insights into miRNA binding in the CDS, we examined the codon adaptation index (CAI) [260] around crosslinked seed matches, and found that the sequence environment of crosslinked seed matches differed from that of noncrosslinked seed matches in the CAI. The bias in codon usage extended for at least 70 codons up- as well as downstream of the crosslinked seed matches (Figure 4.7E), which also correlates well with the marked increase in the A/U content around the binding sites that would lead to a codon usage bias. It was recently reported that miRNA regulation in the CDS was enhanced by inserting rare codons upstream of the miRNA-binding site, presumably due to increased lifetime of miRNA-target-RNA interactions as ribosomes are stalled [261]. These observations suggest that transcripts with reduced translational efficiency form at least transient miRNP complexes amenable to UV crosslinking.

The abundance of mRNAs expressed in HEK293 cells varied over 5 orders of magnitude as shown by DGE profiling. When we related the expression level of CCR-containing transcripts with the magnitude of transcript stabilization after miRNA inhibition, we found that miRNAs preferentially act on transcripts with low and medium expression levels (Figure 4.7F). Highly expressed mRNAs appear to avoid miRNA regulation [190], at least for those miRNAs expressed in HEK293 cells. However, we cannot fully rule out that the weaker response of highly abundant targets may be due to lower affinity and reduced occupancy of miRNA binding sites in highly abundant transcripts.

Earlier studies defining miRNA target regulation were carried out by transfection of miRNAs into cellular systems originally devoid of these miRNAs [81, 84, 85]. We transfected miRNA duplexes corresponding to the deeply conserved miR-7 and miR-124 into FLAG/HA-AGO2 cells, performed PAR-CLIP, and also recorded the effect on mRNA stability upon miR-7 and miR-124 transfection by microarray analysis. Transcripts containing miR-7- or miR-124-specific CCRs were destabilized, especially when CCRs were located in the 3' UTR.

4.2.11 Context Dependence of miRNA Binding

Not every seed-complementary sequence in the HEK293 transcriptome yielded a CCR, thereby providing an opportunity to identify sequence context features specifically contributing to miRNA target binding and crosslinking. For seed-complementary sites that were crosslinked and those that were not crosslinked, we computed the evolutionary selection pressure by the ElMMo method [68], the mRNA stability scores by TargetScan context score [69], and sequence composition and structure measures for the regions around the miRNA seed complementary sites. The feature that distinguished most crosslinked from noncrosslinked seed matches was a 25% lower free energy required to resolve local secondary structure involving the miRNA-binding region, as-

sociated with a 6% increase in the A/U content within 100 nt around the seed-pairing site. These differences were similar for sites located in the CDS and 3' UTRs. Compared to noncrosslinked sites, crosslinked sites are under stronger evolutionary selection (ElMMo) and in sequence contexts facilitating miRNA-dependent mRNA degradation (TargetScan context score).

The location of AGO CCRs within transcript regions was nonrandom and 7-mer or 8-mer sites within the 3' UTR were preferentially located near the stop codon or the polyA tail in transcripts with relatively long 3' UTRs (more than 3 kb). The location of CCRs in the CDS was biased toward the stop codon for the transfected miR-7 and 124, but not for the endogenous miRNAs.

Finally, we wanted to examine how miRNA targets defined by PAR-CLIP compared in regulation of target mRNA stability to those predicted by ElMMo [68], TargetScan context score [69], TargetScan Pct [65] and PicTar [262]. In each case, we selected the same number of highest-scoring sites containing a 7-mer seed-complement to the top 5 expressed miRNAs (let-7a, miR-103, miR-15a, miR-19a, and miR-20a). The analysis was limited to 3' UTR sites due to restriction by the prediction methods. The effect on mRNA stability, as assessed by miRNA antisense inhibition, was overall equivalent for transcripts harboring CCRs compared to transcripts predicted by ElMMo, TargetScan context score, TargetScan Pct and PicTar.

4.3 Discussion

Maturation, localization, decay and translational regulation of mRNAs involve formation of complexes of RBPs and RNPs with their RNA targets [218, 219]. Several hundred RBPs are encoded in the human genome, many of them containing combinations of RNA-binding domains which are drawn from a relatively small repertoire, resulting in diverse structural arrangements and different specificities of target RNA recognition [263]. Furthermore hundreds of miRNAs function together with AGO and TNRC6 proteins to destabilize target mRNAs and/or repress their translation [67]. Collectively, these factors and their presumably combinatorial action constitute the code for posttranscriptional gene regulation. Here we describe an approach to directly identify transcriptome-wide mRNA-binding sites of regulatory RBPs and RNPs in live cells.

4.3.1 PAR-CLIP Allows High-Resolution Mapping of RBP and miRNA Target Sites

We showed that application of photoactivatable nucleoside analogs to live cells facilitates RNA-protein crosslinking and transcriptome-wide identification of RBP and RNP binding sites. We concentrated on 4SU after it became apparent that the crosslinking

sites in isolated RNAs were revealed upon sequencing by a prominent transition from T to C in the cDNA prepared from the isolated RNA segments. Compared to regular UV 254 nm crosslinking in the absence of photoactivatable nucleosides, our method has two distinct advantages. We obtain higher yields of crosslinked RNAs using similar radiation intensities, and more importantly, we can identify crosslinked regions by mutational analysis. Studies using conventional UV 254 nm CLIP have not reported the incidence of deletions and substitutions [230,231,236,237], except for recent work by Grannemann et al. on the U3 snoRNA that showed an increase of deletions at the RBP binding site [234]. Our own analysis indicates that mutations in sequence reads derived from UV 254 nm CLIP were at least one order of magnitude less frequent than T to C transitions observed in PAR-CLIP.

From an experimental perspective, it is important to note that crosslinked RNA segments, irrespective of the methods of isolation, are always contaminated with noncrosslinked RNAs, as shown by consistent identification of rRNAs, tRNAs, and miRNAs. Compared to crosslinked RNA fragments, these unmodified RNA molecules are more readily reverse transcribed, which underscores the need for separation of crosslinked signal from noncrosslinked noise. We now provide a method that accomplishes this critical task.

4.3.2 Context Dependence of 4SU Crosslink Sites

It is conceivable that binding sites located in peculiar sequence environments, e.g., those completely devoid of U, may exist and cannot be captured using 4SU-based crosslinking. However, such sites are extremely rare. Only about 0.4% of 32-nt long sequence segments, representative of the length of our Solexa sequence reads, are U-less, corresponding to an occurrence of one such segment in every 8 kb of a transcript.

Nonetheless, to provide a means to resolve such unlikely situations, we explored the use of other photoactivatable nucleosides, such as 6SG to identify IGF2BP1 binding sites. We found a good correlation between the sequence reads obtained from a given gene with 4SU and 6SG (Pearson correlation coefficient 0.65). Moreover, the sequence read clusters, representing individual binding sites, overlapped strongly: 59% out of the 47,050 6SG clusters were also identified with 4SU, despite of the fact that the environment of IGF2BP1 binding sites was strongly depleted for guanosine. Interestingly, the sequence reads obtained after 6SG crosslinking were enriched for G to A transitions, pointing to a structural change in 6SG analogous to the situation in PAR-CLIP with 4SU. Because 6SG appears to have lower crosslinking efficiency compared to 4SU, we recommend to first use 4SU and then resort to 6SG when the data indicates that the sites of interest are located in sequence contexts devoid of uridines. It is important to point out that neither of these photoactivatable nucleotides appears to be toxic under our recommended conditions.

4.3.3 miRNA Target Identification

When applying PAR-CLIP to isolate miRNA-binding sites, we were surprised to find nearly 50% of the binding sites located in the CDS. However, miRNA inhibition experiments showed that miRNA binding at these sites only caused small, yet significant mRNA destabilization. In spite of the difference in their efficiency of triggering mRNA degradation, CDS and 3' UTR sites appear to have similar sequence and structure features. The sequence bias around CDS sites is associated with an increased incidence of rare codon usage, which could in principle reduce translational rate, thereby providing an opportunity for transient miRNP binding and regulation. Similar observations were made previously using artificially designed reporter systems [261].

The use of the knowledge of the crosslinking site allowed us to narrowly define the miRNA-binding regions for matching the site with the most likely miRNA endogenously co-expressed with its targets, and to assess noncanonical miRNA-binding modes. We were able to explain the majority of PAR-CLIP binding sites by conventional miRNA-mRNA seed-pairing interactions [69], yet found that about 6% of miRNA target sites might best be explained by accepting bulges or mismatches in the seed pairing region, similar to the interaction between let-7 and its target lin-41 [264] and those recently observed in biochemical and structural studies of *T. thermophilus* Ago protein [29, 253].

4.3.4 The mRNA Ribonucleoprotein Code and Its Impact on Gene Regulation

We were able to identify all of the crosslinkable RNA-binding sites present in about 9,000 of the top-expressed mRNA in HEK293 cells representing approximately 95% of the total mRNA molecules of a cell. One of the surprising outcomes of our study was that each of the examined RBPs or miRNPs bound and presumably controlled between 5 and 30% of the more than 20,000 transcripts detectable in HEK293 cells. These results demonstrate that a transcript will generally be bound and regulated by multiple RBPs, the combination of which will determine the final gene-specific regulatory outcome. Exhaustive highresolution mapping of RBP– and RNP–target-RNA interactions is critical, because it may lead to the discovery of specific combination of sites (or modules) that may control distinct cellular processes and pathways. To gain further insights into the dynamics of mRNPs it will be important to also map the sites of RNA-binding factors, such as helicases, nucleases or polymerases, where the specificity determinants are poorly understood. The precise identification of RNA interaction sites will be extremely useful for interrogating the rapidly emerging data on genetic variation between individuals and whether some of these variations possibly contribute to complex genetic diseases by affecting posttranscriptional gene regulation.

4.4 Experimental Procedures

4.4.1 PAR-CLIP

Human embryonic kidney (HEK) 293 cells stably expressing FLAG/HA-tagged IGF2BP1-3, QKI, PUM2, AGO1-4, and TNRC6A-C [192] were grown overnight in medium supplemented with 100 μM 4SU. Living cells were irradiated with 365 nm UV light. Cells were harvested and lysed in NP40 lysis buffer. The cleared cell lysates were treated with RNase T1. FLAG/HA-tagged proteins were immunoprecipitated with anti-FLAG antibodies bound to Protein G Dynabeads. RNase T1 was added to the immunoprecipitate. Beads were washed and resuspended in dephosphorylation buffer. Calf intestinal alkaline phosphatase was added to dephosphorylate the RNA. Beads were washed and incubated with polynucleotide kinase and radioactive ATP to label the crosslinked RNA. The protein-RNA complexes were separated by SDS-PAGE and electroeluted. The electroeluate was proteinase K digested. The RNA was recovered by acidic phenol/chloroform extraction and ethanol precipitation. The recovered RNA was turned into a cDNA library as described [242] and Solexa sequenced. The extracted sequence reads were mapped to the human genome (hg18), human mRNAs and miRNA precursor regions. For a more detailed description of the methods, see the Extended Experimental Procedures. For a video presenting the procedure please visit http://www.jove.com/index/Details.stp?ID=2034.

4.4.2 Oligonucleotide Transfection and mRNA Array Analysis

siRNA, miRNA and 20-O-methyl oligonucleotide transfections of HEK293 T-REx Flp-In cells were performed in 6-well format using Lipofectamine RNAiMAX (Invitrogen) as described by the manufacturer. Total RNA of transfected cells was extracted using TRIZOL following the instructions of the manufacturer. The RNA was further purified using the RNeasy purification kit (QIAGEN). 2 μg of purified total RNA was used in the One-Cycle Eukaryotic Target Labeling Assay (Affymetrix) according to manufacturer's protocol. Biotinylated cRNA targets were cleaned up, fragmented, and hybridized to Human Genome U133 Plus 2.0 Array (Affymetrix). For details of the analysis, see Bioinformatics section in the Supplementary Material.

4.4.3 Generation of Digital Gene Expression (DGEX) Libraries

1 μg each of total RNA from HEK293 cells inducibly expressing tagged IGF2BP1 before and after induction was converted into cDNA libraries for expression profiling by sequencing using the DpnII DGE kit (Illumina) according to instructions of the manufacturer. For details of the analysis, see Bioinformatics section in the Supplemental Information.

4.5 Acknowledgments

We thank V. Hovestadt for his help with the analysis of the crosslinking positions within miRNAs. We are grateful to W. Zhang and C. Zhao (Genomics Resource Center) for mRNA array analysis and Solexa sequencing. We thank Millipore for the antibodies. We thank members of the Tuschl laboratory for comments on the manuscript. M.H. is supported by the Deutscher Akademischer Austauschdienst (DAAD). This work was supported by the Swiss National Fund Grant #3100A0-114001 to M.Z.; T.T. is an HHMI investigator, and work in his laboratory was supported by NIH grants GM073047 and MH08442 and the Starr Foundation.

T.T. is a cofounder and scientific advisor to Alnylam Pharmaceuticals and an advisor to Regulus Therapeutics.

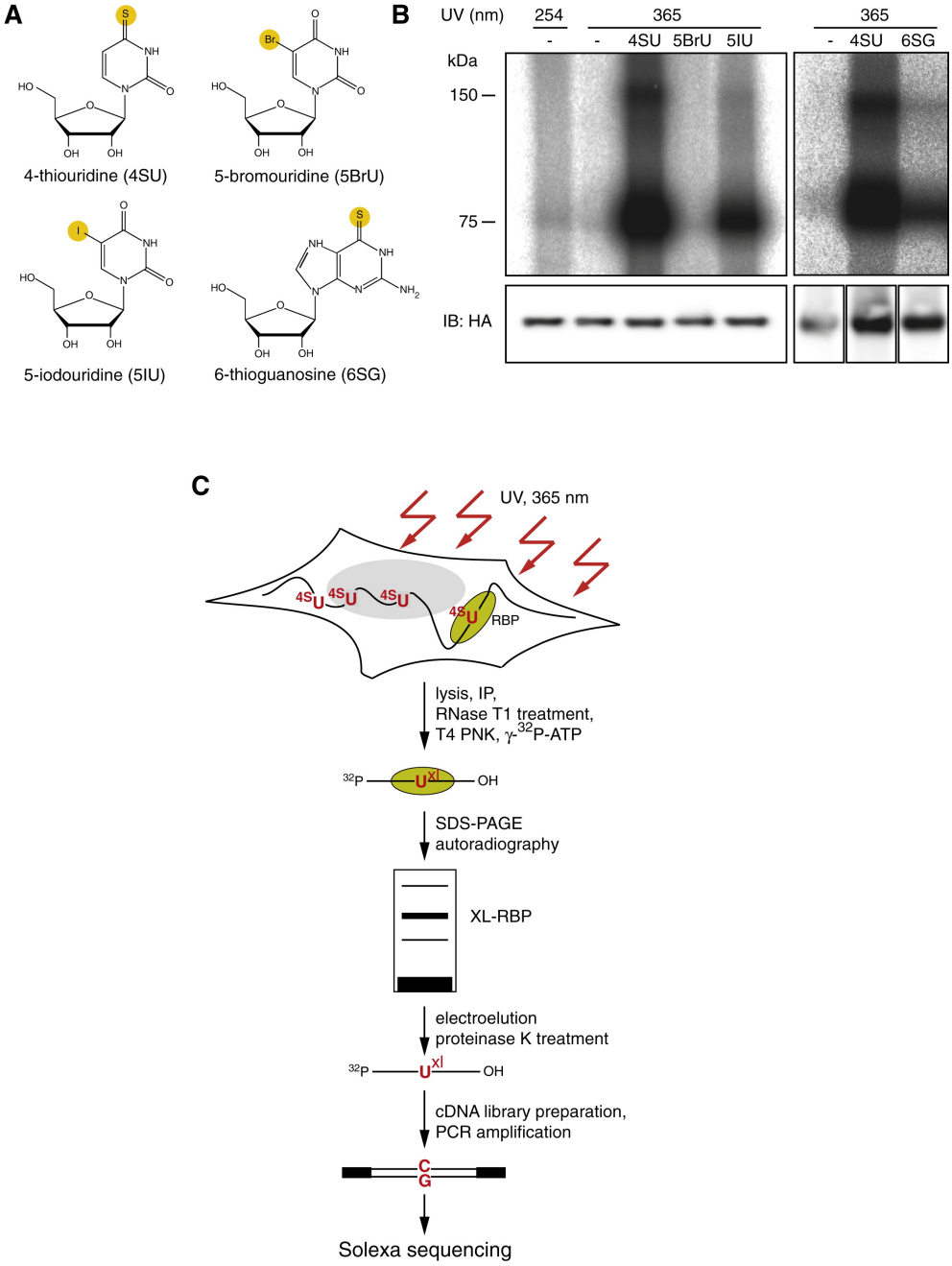


Figure 4.1: **PAR-CLIP Methodology.** (**A**) Structure of photoactivatable nucleosides. (**B**) Phosphorimages of SDS-gels that resolved 5'-^{32}P-labeled RNA–FLAG/HA-IGF2BP1 immunoprecipitates (IPs) prepared from lysates from cells that were cultured in media in the absence or presence of 100 μM photoactivatable nucleoside and crosslinked with UV 365 nm. For comparison, a sample prepared from cells crosslinked with UV 254 nm, was included. Lower panels show immunoblots probed with an anti-HA antibody. (**C**) Illustration of PAR-CLIP. 4SU-labeled transcripts were crosslinked to RBPs and partially RNase-digested RNA-protein complexes were immunopurified and size-fractionated. RNA molecules were recovered and converted into a cDNA library and deep sequenced.

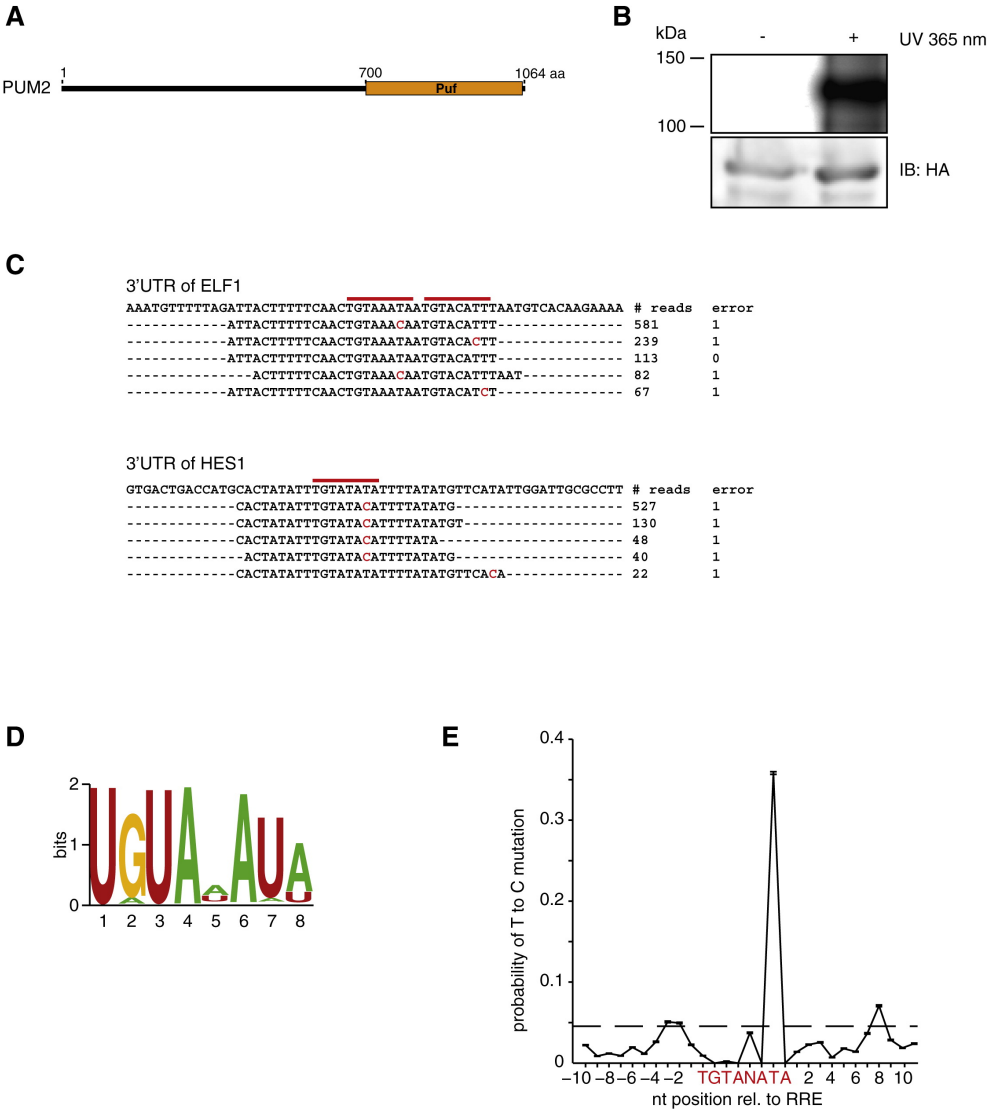


Figure 4.2: **RNA Recognition by PUM2 Protein.** (**A**) Domain structure of PUM2 protein. (**B**) Phosphorimage of SDS-gel of radiolabeled FLAG/HA-PUM2-RNA complexes from nonirradiated or UV-irradiated 4SU-labeled cells. The lower panel shows an anti-HA immunoblot. (**C**) Alignments of PAR-CLIP cDNA sequence reads to corresponding regions in the 3' UTR of ELF1 and HES1 Refseq transcripts. The number of sequence reads (# reads) and mismatches (errors) are indicated. Red bars indicate the PUM2 recognition motif and red-letter nucleotides indicate T to C sequence changes. (**D**) Sequence logo of the PUM2 recognition motif generated by PhyloGibbs analysis of the top 100 sequence read clusters. (**E**) T to C positional mutation frequency for PAR-CLIP clusters anchored at the 8-nt recognition motif from all motif-containing clusters. The dashed line represents the average T to C mutation frequency within these clusters.

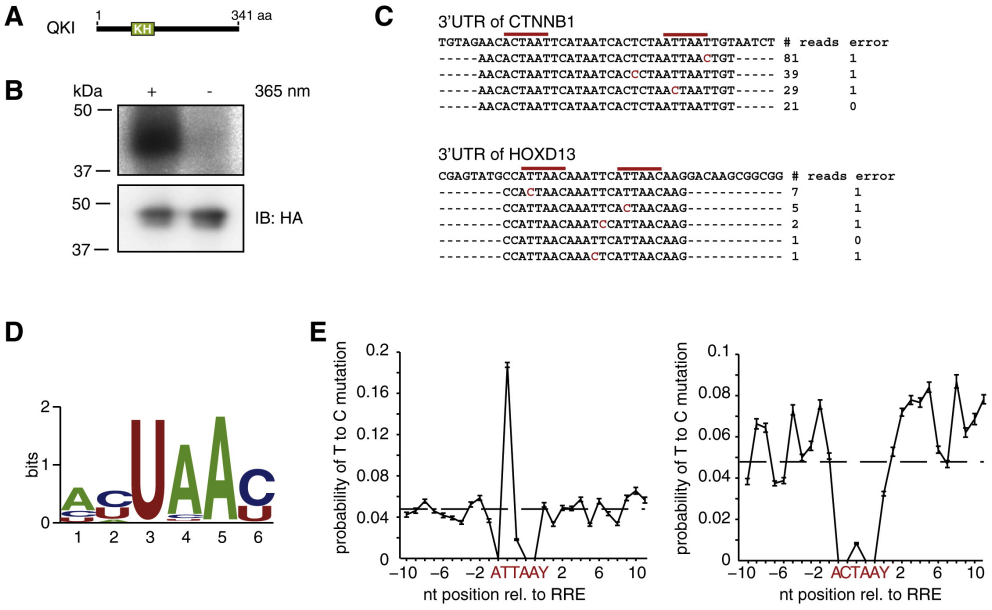


Figure 4.3: **RNA Recognition by QKI Protein.** (**A**) Domain structure of QKI protein. (**B**) Phosphorimage of SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-QKI IPs from nonirradiated or UV-irradiated 4SU-labeled cells. The lower panel shows the anti-HA immunoblot. (**C**) Alignments of PAR-CLIP cDNA sequence reads to the corresponding regions in the 3' UTRs of the CTNNB1 and HOXD13 transcripts. Red bars indicate the QKI recognition motif and red-letter nucleotides indicate T to C sequence changes. (**D**) Sequence logo of the QKI recognition motif generated by PhyloGibbs analysis of the top 100 sequence read clusters. (**E**) T to C positional mutation frequency for PAR-CLIP clusters anchored at the AUUAAY (left panel) and ACUAAY (right panel) RRE; Y = U or C. The dashed line represents the average T to C mutation frequency within these clusters.

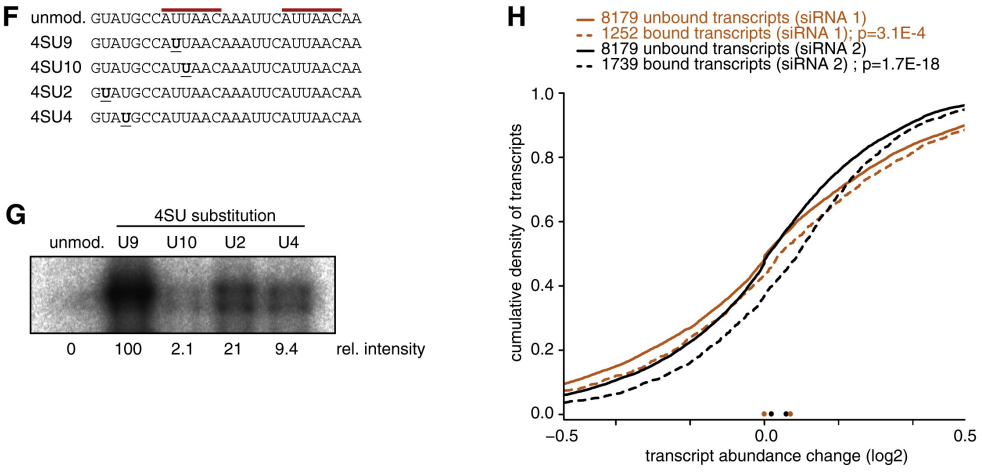


Figure 4.3: **RNA Recognition by QKI Protein.** (**F**) Sequences of synthetic 4SU-labeled oligoribonucleotides with QKI recognition motifs, derived from a sequence read cluster aligning to the 3' UTR of HOXD13 shown in (**C**) 4SU-modified residues are underlined. (**G**) Phosphorimage of SDS-gel resolving recombinant QKI protein after crosslinking to radiolabeled synthetic oligoribonucleotides shown in (**F**). (**H**) Stabilization of QKI-bound transcripts upon siRNA knockdown. Two distinct siRNA duplexes (1, orange traces and 2, black traces) were used for QKI knockdown and changes in transcript stability relative to mock transfection were inferred from microarray analysis. Shown are the distributions of changes upon siRNA transfection for transcripts that did (dashed lines) or did not (solid lines) contain QKI PAR-CLIP clusters. The p-values obtained in the Wilcoxon rank-sum test comparing the changes in targeted and nontargeted transcripts are indicated.

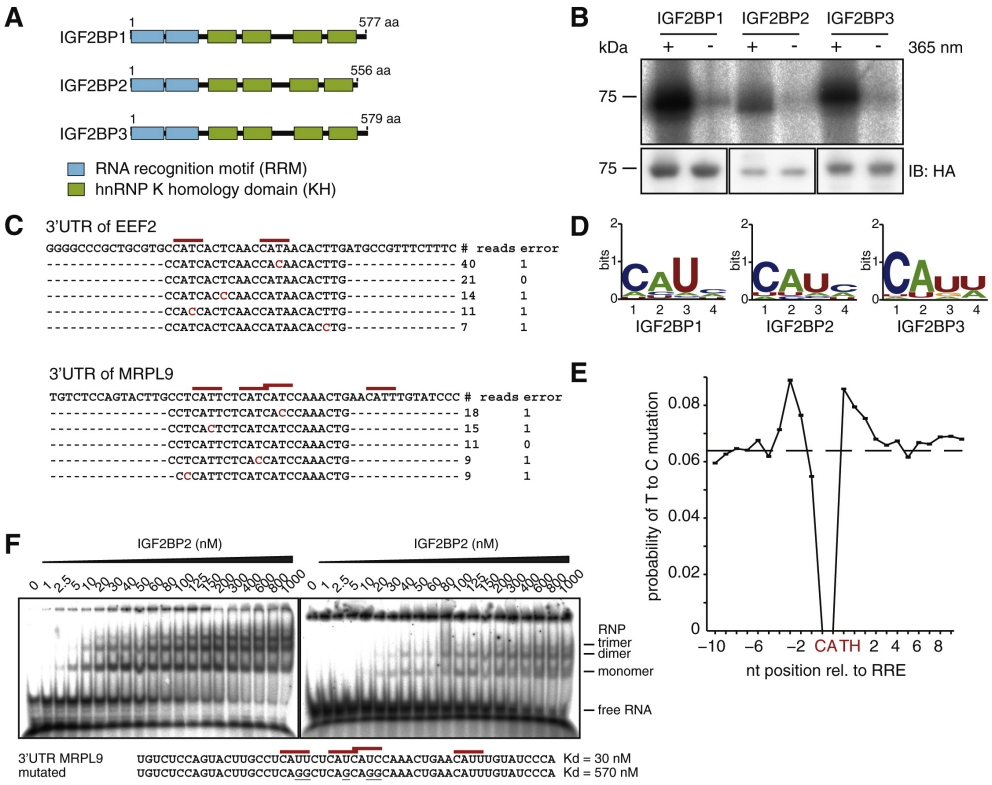


Figure 4.4: **RNA Recognition by the IGF2BP Protein Family.** (**A**) Domain structure of IGF2BP1-3 proteins. (**B**) Phosphorimage of an SDS-gel resolving radiolabeled RNA crosslinked to FLAG/HA-IGF2BP1-3 IPs. The lower panel shows anti-HA immunoblots. (**C**) Alignments of IGF2BP1 PAR-CLIP cDNA sequence reads to the corresponding regions of the 3' UTRs of EEF2 and MRPL9 transcripts. Red bars indicate the 4-nt IGF2BP1 recognition motif and nucleotides marked in red indicate T to C sequence changes. (**D**) Sequence logo of the IGF2BP1-3 RRE generated by PhyloGibbs analysis of the top 100 sequence read clusters. (**E**) T to C positional mutation frequency for PAR-CLIP clusters anchored at the 4-nt recognition motif from all motif-containing clusters. The dashed line represents the average T to C mutation frequency within these clusters. (**F**) Phosphorimage of native PAGE resolving complexes of recombinant IGF2BP2 protein with wild-type (left panel) and mutated target oligoribonucleotide (right panel). Sequences and dissociation constants (Kd) are indicated.

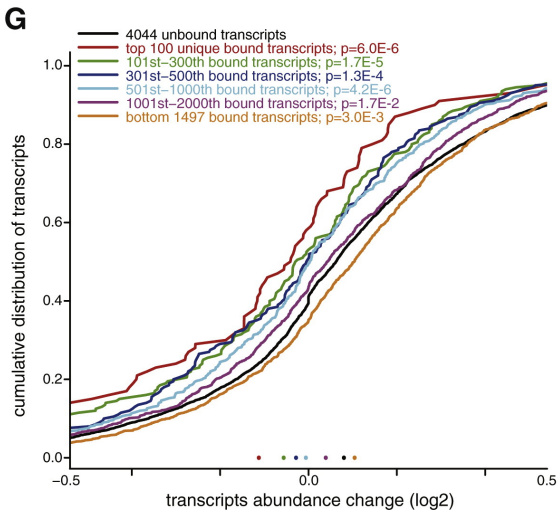


Figure 4.4: **RNA Recognition by the IGF2BP Protein Family.** (**G**) Destabilization of IGF2BP-bound transcripts upon siRNA knockdown. A cocktail of three siRNA duplexes targeting IGF2BP1, 2, and 3 was used, as well as a mock transfection and changes in transcript stability were monitored by microarray analysis. Distributions of transcript level changes for IGF2BP1-3 PAR-CLIP target transcripts versus nontargeted transcripts are shown. IGF2BP1-3 target sequences were ranked and divided into bins. The p-values indicate the significance of the difference between the changes of target versus nontarget transcripts, as given by the Wilcoxon rank-sum test and are corrected for multiple testing.

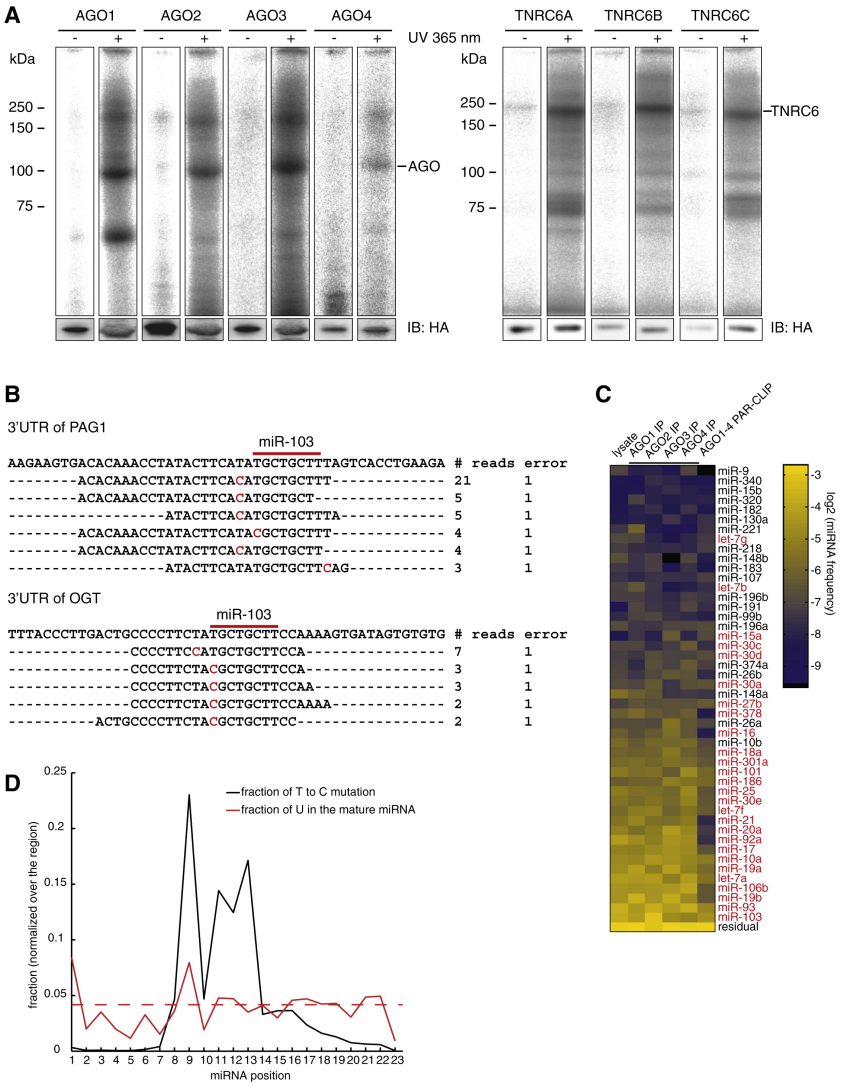


Figure 4.5: **AGO Protein Family and TNRC6 Family PAR-CLIP.** (**A**) Phosphorimage of SDS-gels resolving radiolabeled RNA crosslinked to the FLAG/HA-AGO1-4 and FLAG/HA-TNRC6A-C IPs. The lower panel shows the immunoblot with an anti-HA antibody. (**B**) Alignment of AGO PAR-CLIP cDNA sequence reads to the corresponding regions of the 3' UTRs of PAG1 and OGT. Red bars indicate the 8-nt miR-103 seed complementary sequence and nucleotides marked in red indicate T to C mutations. (**C**) miRNA profiles from RNA isolated from untreated HEK293 cells, noncrosslinked FLAG/HA-AGO1-4 IPs, and combined AGO1-4 PAR-CLIP libraries. The color code represents relative frequencies determined by sequencing. miRNAs indicated in red were inhibited by antisense oligonucleotides for the transcriptome-wide characterization of the destabilization effect of miRNA binding. (**D**) T to C positional mutation frequency for miRNA sequence reads is shown in black, and the normalized frequency of occurrence of uridines within miRNAs is shown in red. The dashed red line represents the normalized mean U frequency in miRNAs.

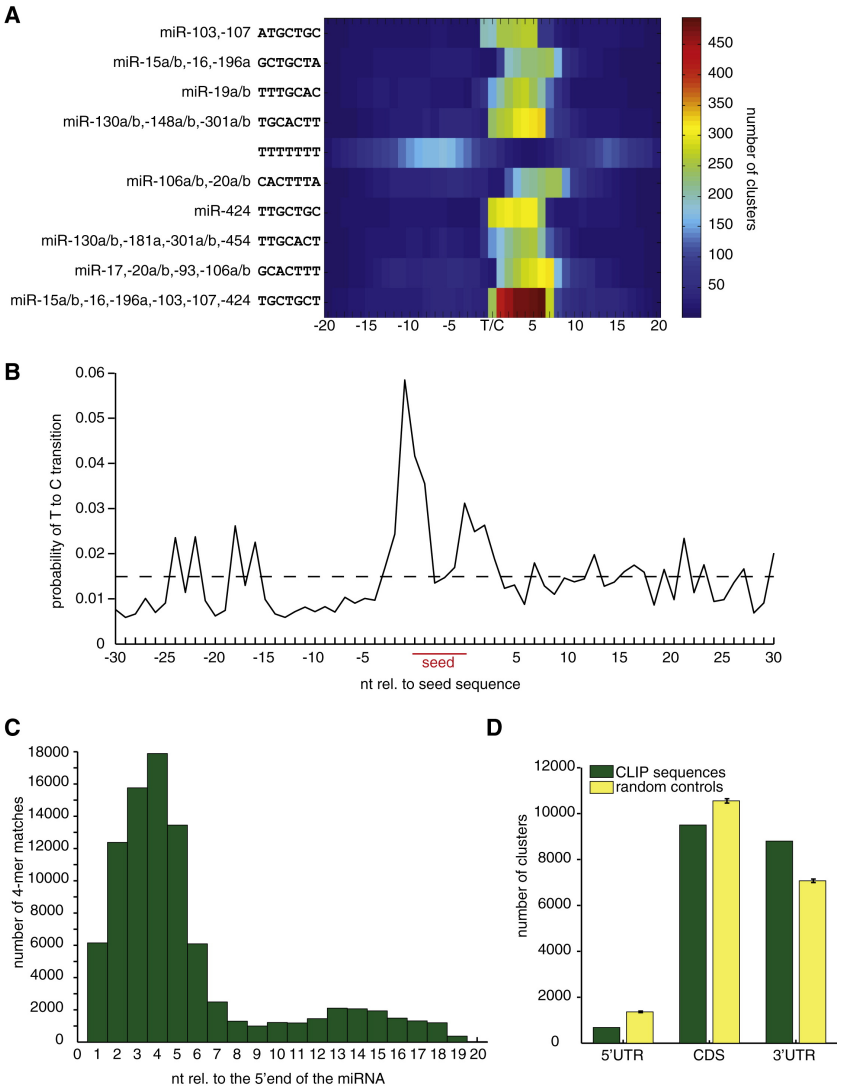


Figure 4.6: **AGO PAR-CLIP Identifies miRNA Seed-Complementary Sequences in HEK293 Cells** (**A**) Representation of the 10 most significantly enriched 7-mer sequences within PAR-CLIP CCRs. T/C indicates the predominant T to C transition within clusters of sequence reads. (**B**) T to C positional mutation frequency for clusters of sequence reads anchored at the 7-mer seed complementary sequence (pos. 2–8 of the miRNA) from all clusters containing seed-complementary sequences to any of the top 100 expressed miRNAs in HEK293 cells. The dashed line represents the average T to C mutation frequency within the clusters. (**C**) Identification of 4-nt base-pairing regions contributing to miRNA target recognition. CCRs with at least one 7-mer seed complementary region to one of the top 100 expressed miRNAs were selected. The number of 4-nt contiguous matches in the CCRs relative to the 5' end of the matching miRNA was counted. (**D**) Analysis of the positional distribution of CCRs. The number of clusters annotated as derived from the 5' UTR, CDS or 3' UTR of target transcripts is shown (green bars). Yellow bars show the expected location distribution of the crosslinked regions if the AGO proteins bound without regional preference to the target transcript.

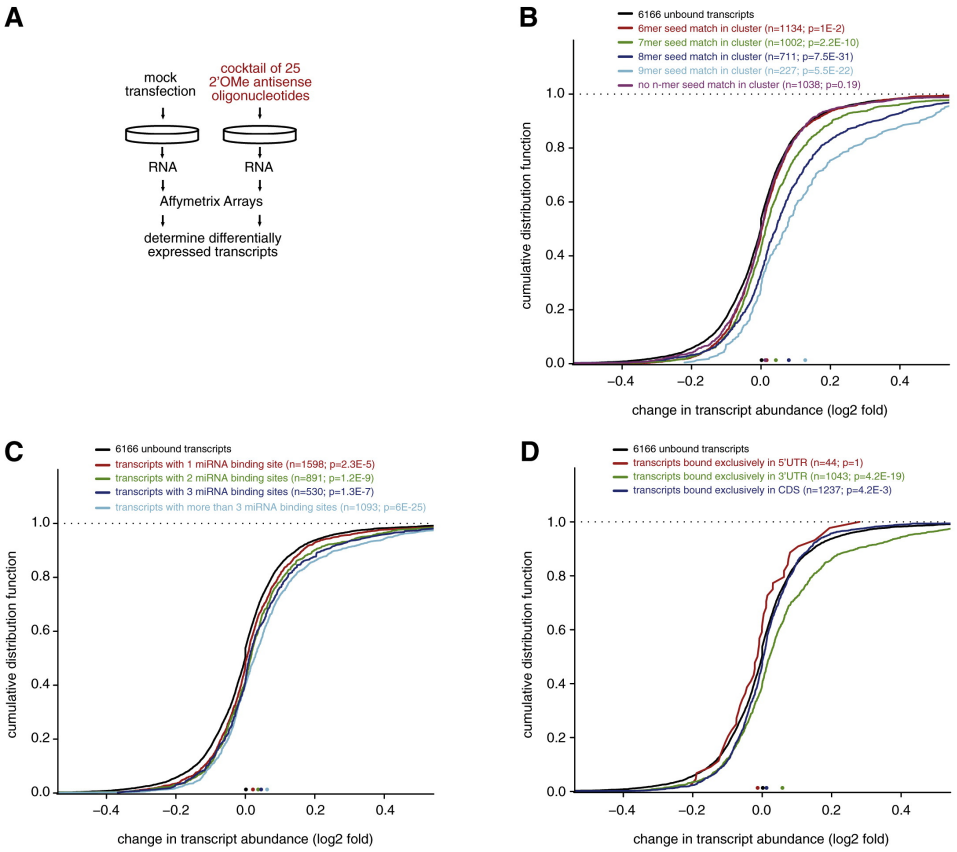


Figure 4.7: **Relationship between Various Features of miRNA/Target RNA Interactions and mRNA Stability.** (**A**) FLAG/HA-AGO2-tagged HEK293 cells were transfected with a cocktail of 25 2'-O-methyl modified antisense oligoribonucleotides, inhibiting miRNAs marked in red in Figure 4.5C, or mock transfected, followed by microarray analysis of the change of mRNA expression levels. (**B**) Transcripts containing CCRs were categorized according to the presence of n-mer seed complementary matches and the distributions of stability changes upon miRNA inhibition are shown for these categories. The stability change for transcripts harboring CCRs without identifiable miRNA seed-complementary regions is also shown. The p values indicate the significance of the difference between the transcript level changes of transcripts containing CCRs versus transcripts without CCRs, as given by the Wilcoxon rank-sum test and are corrected for multiple testing. (**C**) Transcripts were categorized according to the number of CCRs they contained. (**D**) Transcripts were categorized according to the positional distribution of CCRs. Only transcripts containing CCRs exclusively in the indicated region are used.

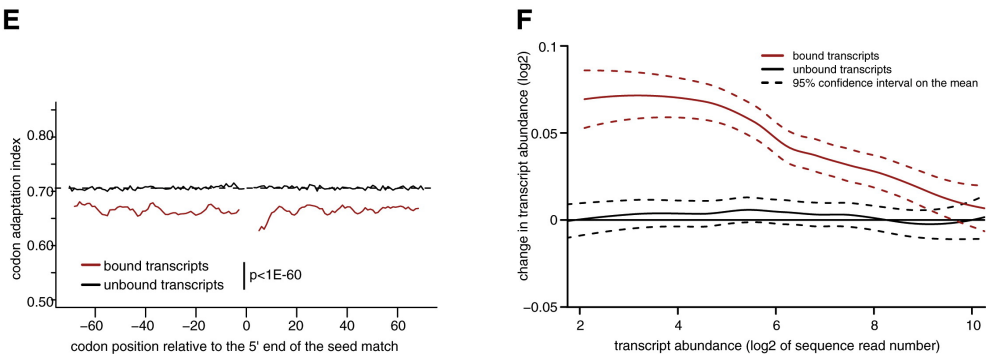


Figure 4.7: **Relationship between Various Features of miRNA/Target RNA Interactions and mRNA Stability.** (**E**) Codon adaptation index (CAI) for transcripts containing 7-mer seed complementary regions (pos. 2-8) in the CDS for the miR-15, miR-19, miR-20, and let-7 miRNA families. The red and the black lines indicate the CAI for seed-complementary sequence containing transcripts bound and not bound by AGO proteins determined by AGO PAR-CLIP. (**F**) LOESS regression of total transcript abundance in HEK293 cells ($\log_2$ of sequence counts determined by digital gene expression (DGE)) against fold change of transcript abundance ($\log_2$) determined by microarrays after transfection of the miRNA antagonist cocktail versus mock transfection of AGO-bound and unbound transcripts.

Chapter 5

MicroRNAs control *de novo* DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells

Lasse Sinkkonen[1], Tabea Hugenschmidt[1], Philipp Berninger[2], Dimos Gaidatzis[2], Fabio Mohn[1], Caroline G Artus-Revel[1], Mihaela Zavolan[2], Petr Svoboda[3,4], and Witold Filipowicz[1,4]

1 Friedrich Miescher Institute for Biomedical Research, Maulbeerstrasse 66, 4058 Basel, Switzerland.

2 Division of Bioinformatics, Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland, and the Swiss Institute of Bioinformatics.

3 Institute of Molecular Genetics, Videnska 1083, Prague, Czech Republic.

4 corresponding authors

published in *Nature Structural & Molecular Biology 15, 259-267, 2008*

Loss of microRNA (miRNA) pathway components negatively affects differentiation of embryonic stem (ES) cells, but the underlying molecular mechanisms remain poorly defined. Here we characterize changes in mouse ES cells lacking Dicer (Dicer1). Transcriptome analysis of *Dicer*$^{-/-}$ cells indicates that the ES-specific miR-290 cluster has an important regulatory function in undifferentiated ES cells. Consistently, many of the

defects in Dicer-deficient cells can be reversed by transfection with miR-290 family miRNAs. We demonstrate that *Oct4* (also known as *Pou5f1*) silencing in differentiating *Dicer*$^{-/-}$ ES cells is accompanied by accumulation of repressive histone marks but not by DNA methylation, which prevents the stable repression of *Oct4*. The methylation defect correlates with downregulation of *de novo* DNA methyltransferases (Dnmts). The downregulation is mediated by Rbl2 and possibly other transcriptional repressors, potential direct targets of miR-290 cluster miRNAs. The defective DNA methylation can be rescued by ectopic expression of *de novo* Dnmts or by transfection of the miR-290 cluster miRNAs, indicating that *de novo* DNA methylation in ES cells is controlled by miRNAs.

5.1 Introduction

Short 20-25-nucleotide (nt) RNAs have emerged recently as important sequence-specific regulators of gene expression in eukaryotes [178, 265–267]. Short RNAs are produced from long double-stranded RNA (dsRNA) and miRNA precursors, which are processed by the RNase III family enzymes Drosha and Dicer to yield mature effector molecules, small interfering RNAs (siRNAs) and miRNAs [178, 265–268]. MiRNAs are the dominant class of short RNAs in mammalian cells, from which several hundred different miRNAs have been identified and implicated in the regulation of many cellular processes [41, 269]. Mammalian miRNAs typically base-pair imperfectly with the 3' untranslated region (3' UTR) of target mRNAs and induce their translational repression or degradation [270, 271]. The eight 5' terminal nucleotides form the critical miRNA region for target mRNA recognition. This region, generally referred to as the 'seed', hybridizes nearly perfectly with the target to nucleate the miRNA-mRNA interaction [182, 201]. Most computational methods of miRNA target prediction incorporate this constraint [252].

ES cells are pluripotent cells derived from the inner cell mass of blastocysts. Depending on the culture conditions, ES cells can differentiate into various cell types [272]. The Oct4, Sox2 and Nanog transcription factors form a core circuit responsible for the transcriptional control of ES cell renewal and pluripotency [273, 274]. Mouse ES cells contain numerous miRNAs, including a cluster of six miRNAs (miR-290 through miR-295) that share a 5'-proximal AAGUGC motif [177, 275]. The cluster (for brevity referred to as the miR-290 cluster) is specific to ES cells [177]. Its expression increases during preimplantation development [276] and remains high in undifferentiated ES cells, but decreases after ES cell differentiation [177]. Genes and pathways regulated by the miR-290 cluster are unknown.

The loss of Dicer in mouse ES cells results in miRNA depletion [277, 278] and causes differentiation defects *in vivo* and *in vitro* [277]. *Dicer*$^{-/-}$ cells make no con-

tribution to chimeric mice and fail to generate teratomas *in vivo*. In *vitro*, *Dicer*$^{-/-}$ cells form embryoid body (EB)-like structures, but there is little morphological evidence of differentiation. Expression of *Oct4*, a characteristic marker of pluripotent ES cells, is only partially decreased in mutant EBs after day 5 of differentiation, and expression of endodermal and mesodermal markers is not detectable [277]. Similarly, the loss of Dgcr8, a protein required specifically for miRNA maturation, causes partial downregulation of pluripotency markers during retinoic acid (RA)-induced differentiation [60].

In this work, we investigated the molecular mechanisms underlying the inability of *Dicer*$^{-/-}$ ES cells to differentiate. We found that silencing of the *Oct4* pluripotency factor is properly initiated in differentiating *Dicer*$^{-/-}$ ES cells, but it is not followed by *de novo* DNA methylation of the promoter. Consistent with this, we observed that levels of *de novo* DNA methyltransferases are downregulated in *Dicer*$^{-/-}$ cells in an miR-290 cluster-dependent manner. Thus, our data indicate that the *de novo* DNA methylation in differentiating ES cells is regulated by ES-specific miRNAs from the miR-290 cluster.

5.2 Results

5.2.1 Transcriptome analysis of *Dicer*$^{-/-}$ ES cells

To study the roles of miRNAs in gene regulation in ES cells, we profiled the transcriptomes of *Dicer*$^{-/-}$ and *Dicer*$^{+/-}$ ES cells using Affymetrix microarrays. We found a similar number of transcripts that were upregulated (2,551; P-value < 0.001) and downregulated (2,578; P-value < 0.001) upon the loss of Dicer (Figure 5.1a). Analysis of core pluripotency regulators, as well as different differentiation markers, indicated that *Dicer*$^{-/-}$ cells retain characteristics of undifferentiated ES cells.

The binding of miRNAs to the 3' UTR of mRNAs commonly results in degradation of mRNA targets. Numerous studies have reported significant enrichment of sequences complementary to miRNA seeds in 3' UTRs of mRNAs that are upregulated in miRNA knockdowns, or downregulated upon overexpression of miRNAs [77, 81, 279].We searched for sequence motifs that are enriched in the 3' UTRs of transcripts upregulated in the *Dicer*$^{-/-}$ ES cells and that could explain the mRNA expression changes. The three motifs that were most significantly enriched (Figure 5.1b) were all complementary to the seed region of embryonic miRNAs [206]: miR-291a-3p, miR-291b-3p, miR-294 and miR-295 in the case of the first and second motifs (GCACUUU and AGCACUU), and miR-302 in the case of the third motif (GCACUUA). The seed region of miR-302 differs from that of miR-290 cluster members only in the first nucleotide. The enrichment of the GCACUUA motif may imply that miR-302 has an important role in regulating mRNA expression in ES cells. Alternatively, it may indi-

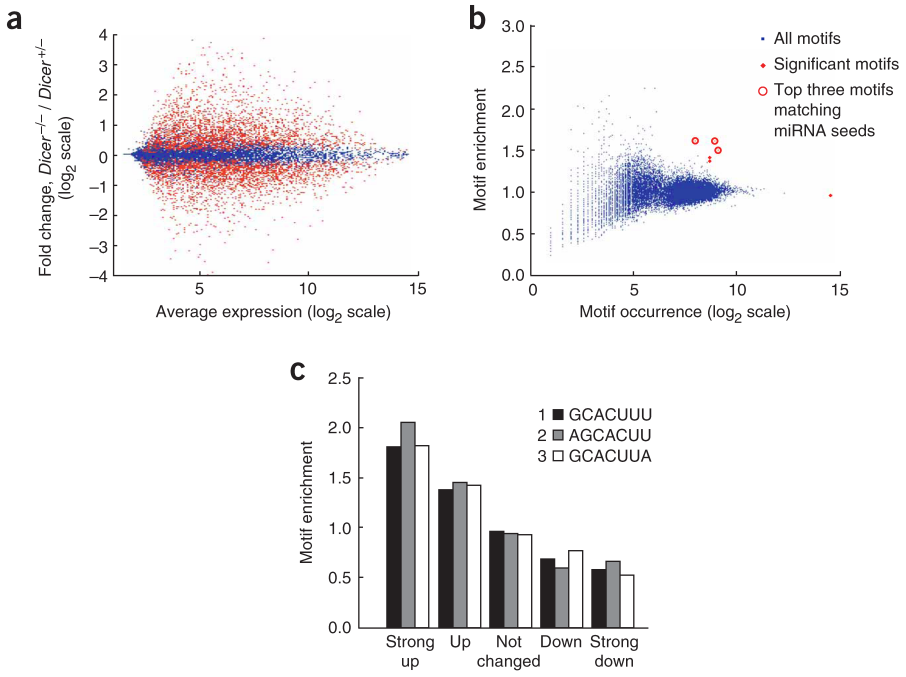


Figure 5.1: **Transcriptome analysis of *Dicer*$^{-/-}$ embryonic stem (ES) cells** (**a**) M [$\log_2$(fold change)] vs. A [average $\log_2$(expression level)] plot for *Dicer*$^{-/-}$ vs. *Dicer*$^{+/-}$ ES cells. Each dot represents a transcript. Significant expression changes (P-value $<$0.001, n = 3) are shown in red. (**b**) Heptamer (7-mer) motif analysis of upregulated transcripts indicates enrichment in motifs complementary to the seed of miR-290 cluster miRNAs. Motifs whose frequency in the 3' UTRs of upregulated transcripts is significantly different from the frequency in the entire set of 3' UTRs are in shown in red. (**c**) Correlation between the occurrence of sequence motifs and the change in mRNA expression. Transcripts were divided into five sets on the basis of their change in expression in *Dicer*$^{-/-}$ compared with *Dicer*$^{-/-}$ ES cells as follows: strong down, more than 2-fold downregulation; down, 1.2-fold to 2-fold downregulation; not changed, 1.2-fold downregulation to 1.2-fold upregulation; up, 1.2-fold to 2-fold upregulation; strong up, more than 2-fold upregulation.

cate that miRNAs prefer target sites with an A residue opposite the 5'-most nucleotide of the miRNA, as has been proposed before [182]. Because the same motif is also most significantly enriched in the 3' UTRs of mRNAs that are downregulated upon transfection with miR-290 cluster miRNAs (see below), we favor the second explanation. We also note that the ubiquitously expressed oncogenic miRNAs of the miR-17/20/93/106 cluster share extensive similarity at their 5' end with the embryonic miRNAs and could also contribute to mRNA regulation in ES cells. As shown in Figure 5.1c, the frequency of the top three motifs decreased gradually from the mRNAs that are most strongly upregulated in *Dicer*$^{-/-}$ cells to the mRNAs that are strongly downregulated.

We examined expression of the miR-290 cluster primary transcript using available microarray data [280]. Quantification of the primary transcript indicated that expression of the cluster occurs zygotically and reaches the highest level in the blastocyst. Notably, accumulation of the miR-290 cluster transcript was downregulated in *Dicer*$^{-/-}$ ES cells, indicating a possible feedback control of its expression by the cluster or other miRNAs. Array analysis of miRNA levels in Dicer$^{+/-}$ and *Dicer*$^{-/-}$ ES cells using Exiqon arrays revealed that, as expected [170,177], miR-290 cluster miRNAs are abundantly expressed in ES cells, and miR-290 cluster and other miRNA levels are reduced in *Dicer*$^{-/-}$ cells.

5.2.2 Identification of primary miR-290 cluster targets

To increase the accuracy of the miRNA target prediction, we compared the transcriptome profile of *Dicer*$^{-/-}$ ES cells (transfected with a nonspecific siRNA as a control) with that of *Dicer*$^{-/-}$ ES cells transfected with the siRNA-like form of miRNAs of the miR-290 cluster (Figure 5.2a). Applying the same heptamer motif analysis used above, we found a few motifs enriched in transcripts that were downregulated after miR-290 cluster miRNA transfection. Among them are motifs complementary to seeds of miR-290 cluster miRNAs, identical to the top three motifs identified above (Figure 5.2b). Analysis of both array experiments showed a good inverse correlation between transcript-level changes in *Dicer*$^{-/-}$ cells (compared to Dicer$^{+/-}$ cells) and *Dicer*$^{-/-}$ cells transfected with miR-290 cluster miRNAs (compared to control *Dicer*$^{-/-}$ cells) (Figure 5.4a). The correlation holds for mRNAs that carry the miR-290 cluster seed-matching sequences in their 3' UTR, as well as for those that do not (Figure 5.4b,c). The correlation for mRNAs lacking seed-matching sequences anywhere in the transcript was as good as that shown in Figure 5.4c. These data suggest that not only primary miRNA effects, but also many secondary gene-expression changes controlled by miR-290 cluster miRNAs, are reversible in *Dicer*$^{-/-}$ ES cells.

To predict primary miR-290 cluster targets, we used data from both sets of microarray experiments. We intersected the lists of transcripts that showed a significant change (P-value <0.001) in the expected direction in the *Dicer*$^{-/-}$ cells compared to

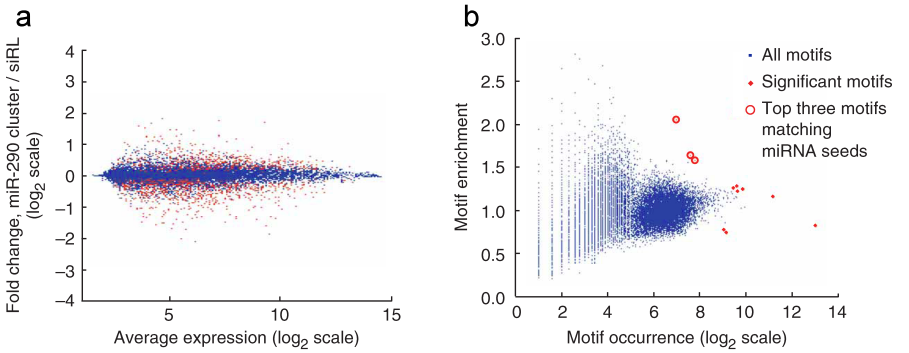


Figure 5.2: **Transcriptome analysis of *Dicer*$^{-/-}$ embryonic stem (ES) cells** (**a**) MA-plot for *Dicer*$^{-/-}$ ES cells transfected with the miR-290 cluster versus *Dicer*$^{-/-}$ ES cells transfected with nonspecific siRNA as control, significant expression changes are shown in red. (**b**) Heptamer motif analysis of downregulated transcripts in the miR-290 cluster-transfected *Dicer*$^{-/-}$ cells. Many significantly enriched motifs are complementary the seeds of the miR-290 cluster miRNAs. The motifs complementary to the seed of siRL did not show any enrichment, indicating that there was a minimal off-target effect.

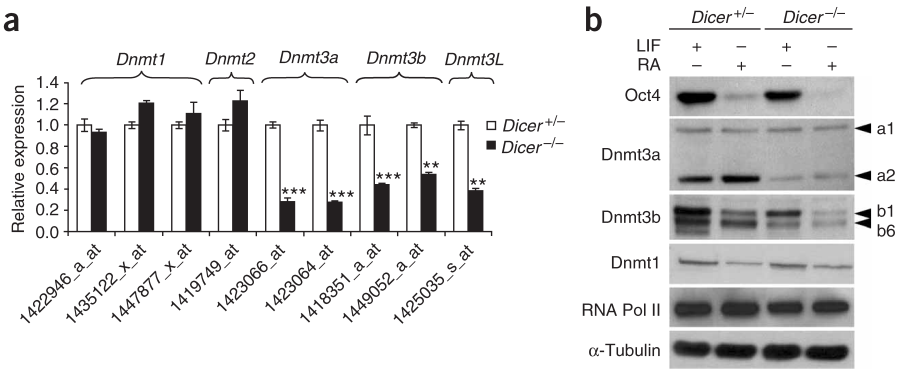


Figure 5.3: ***De novo* DNA methyltransferases (Dnmts) are downregulated in *Dicer*$^{-/-}$ embryonic stem (ES) cells and their expression is rescued by miR-290 cluster miRNAs** (**a**) Expression of DNA methyltransferases in undifferentiated *Dicer*$^{+/-}$ and Dicer- /- cells as analyzed by Affymetrix microarrays. The probe sets detecting mRNAs encoding different DNA methyltransferases are indicated. Mean expression (s.d.; n = 3) in *Dicer*$^{+/-}$ cells was set to one. Signals from probe sets detecting *Dnmt3a*, *Dnmt3b* and *Dnmt3l* were significantly downregulated in *Dicer*$^{-/-}$ cells (two-tailed t-test P-values, from left to right: 0.0001, 0.0006, 0.0093, 0.0022 and 0.0010). (**b**) Western blot analysis of Dnmt1, Dnmt3a and Dnmt3b levels in ES cells cultured in the presence of either leukemia inhibitory factor (LIF) or retinoic acid (RA) for 3 d. α-Tubulin was used as a loading control. Quantification of western blots shown in b and d and in Figure 3f by image densitometry revealed a 3.0-fold to 5.6-fold change in the level of Dnmt3a2 and a 2.0-fold to 4.4-fold change in the levels of Dnmt3b1/b6 between conditions of low and high expression of the proteins.

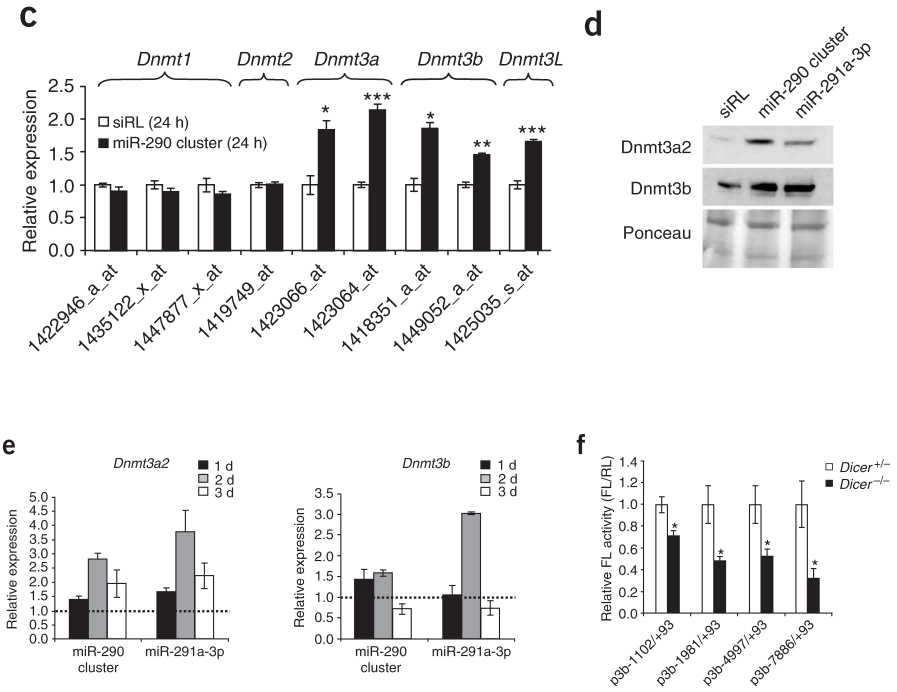


Figure 5.3: ***De novo*** **DNA methyltransferases (Dnmts) are downregulated in *Dicer*$^{-/-}$ embryonic stem (ES) cells and their expression is rescued by miR-290 cluster miRNAs** (**c**) The miR-290 cluster miRNAs induce accumulation of mRNAs encoding Dnmt3a, Dnmt3b and Dnmt3l in *Dicer*$^{-/-}$ ES cells. Mean values (s.d.; n = 3) observed for the siRL-transfected cells (a nonspecific control) were set to one. The P-values, from left to right, were: 0.0102, 0.0008, 0.0021, 0.0010 and 0.0009. (**d**) Dnmt3a2 and Dnmt3b expression 3 d after transfection with siRL, miR-290 cluster or miR-291a-3p. Ponceau staining served as a loading control. (**e**) Upregulation of *Dnmt3a2* and *Dnmt3b1/6* (quantified by RT-qPCR) in response to transfection of either all miR-290 cluster miRNAs or miR-291a-3p into *Dicer*$^{-/-}$ ES cells. Mean expression values (s.d.; n = 3) were normalized to glyceraldehyde-3-phosphate dehydrogenase (Gapdh) and are shown relative to corresponding siRL samples, whose expression values were set to one (dashed line). (**f**) Dicer loss affects transcription from the Dnmt3b promoter. Firefly luciferase (FL) reporters containing Dnmt3b promoter fragments were co-transfected to *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ ES cells together with the pRL-TK control reporter. Mean FL activity values (s.e.m., n 3) in *Dicer*$^{+/-}$ ES cells were set to one. The P-values, from left to right, were: 0.0192, 0.0391, 0.0238 and 0.0230.

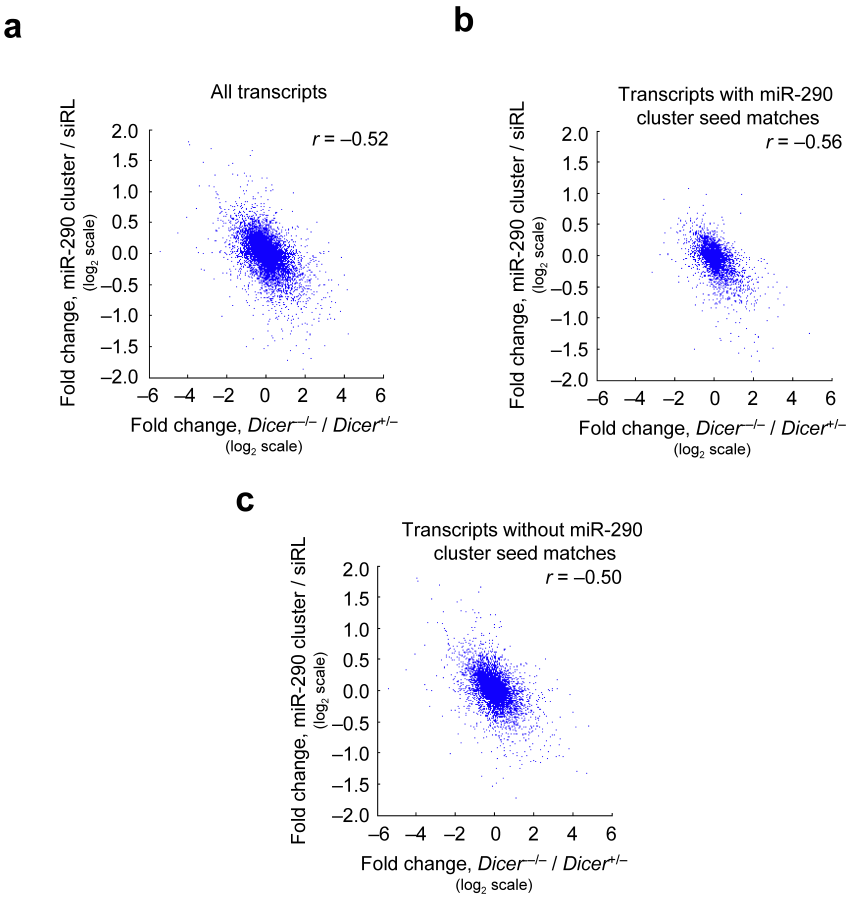


Figure 5.4: **Relationship between the expression change in the *Dicer*$^{-/-}$ vs. Dicer$^{+/-}$ (x-axis) and miR-290s-transfected vs. siRL-transfected *Dicer*$^{-/-}$ ES cells (y-axis)** Each dot corresponds to a single transcript, and the panels represent: (**a**) all transcripts; (**b**) transcripts with at least one 7-mer match to one of the 1-8 positions of the miRNAs of the 290 cluster in their 3' UTRs; (**c**) transcripts with no 7-mer match to any of the miRNAs of the 290 cluster in their 3' UTRs. The spearman rank coefficients are indicated in each panel.

Dicer$^{+/-}$ cells (upregulation) and in the miR-290 cluster-transfected *Dicer*$^{-/-}$ cells compared to control siRNA-transfected *Dicer*$^{-/-}$ cells (downregulation). The list was then filtered to keep only the transcripts whose 3' UTRs had at least one match to the GCACUU hexamer, which is common to all significantly enriched heptamers. The resulting list of predicted targets contained 253 mRNAs. However, it is likely that the number of targets is even larger, as not all expressed mRNAs are detectable by microarrays and some genes may be regulated at the protein rather than the transcript level.

5.2.3 Indirect control of *de novo* methyltransferases by miRNAs

Inspection of microarray data indicated that expression of *de novo* DNA methyltransferase genes *Dnmt3a*, *Dnmt3b* and *Dnmt3l* was significantly downregulated in undifferentiated *Dicer*$^{-/-}$ ES cells (Figure 5.3a). Protein levels of Dnmt3a2, Dnmt3b1 and Dnmt3b6 were also lower in *Dicer*$^{-/-}$ cells, whereas the ubiquitously expressed isoform of Dnmt3a, Dnmt3a1 ([281]), remained unchanged (Figure 5.3b). Notably, expression of *de novo* DNA methyltransferases could be rescued, at both mRNA and protein levels, upon transfection of all miR-290 cluster miRNAs or miR-291a-3p alone (Figure 5.3c-e). Similar downregulation of all three *Dnmt3* genes upon loss of Dicer and their upregulation in response to transfection of miR-290 cluster miRNAs indicated that miRNAs may regulate the expression of *Dnmt3* genes indirectly, possibly by controlling the activity of a common transcriptional repressor. This possibility is supported by the observations that *Dnmt3a2*, *Dnmt3b* and *Dnmt3l* contain similar TATA-less GC-rich promoters, are regulated by SP1-SP3 transcription factors, and are highly expressed in blastocysts and ES cells but are downregulated during differentiation into somatic lineages [281–284]. To corroborate the possibility of transcriptional regulation, we compared the activity of firefly luciferase (FL) reporters containing Dnmt3b promoter regions of different lengths. Activity of the reporters was significantly lower in *Dicer*$^{-/-}$ than in Dicer$^{+/-}$ ES cells (Figure 5.3f), arguing that the *Dnmt3b* promoter is markedly repressed in cells lacking Dicer and suggesting that downregulation of *Dnmt3* genes in *Dicer*$^{-/-}$ ES cells may occur at the level of transcription.

Among the predicted primary targets of the miR-290 cluster, we identified several annotated [285] transcriptional repressors that are upregulated during embryonic differentiation after the blastocyst stage [284]. They include genes for the basic Kruppel-like factor *Klf3*, the nuclear receptor *Nr2f2*, the zinc-finger proteins *Zmynd11* and *Zbtb7*, and retinoblastoma-like 2 (*Rbl2*) (Figure 5.5a,b). Several other observations make *Rbl2* a plausible candidate for the miR-290 cluster-regulated transcriptional repressor of *de novo* DNA methyltransferases. The *Rbl2* 3' UTR contains conserved potential binding sites for miR-290 cluster miRNAs (Figure 5.5c), and *Rbl2* mRNA is downregulated upon transfection of all miR-290 cluster miRNAs or miR-291a-3p alone into

Dicer$^{-/-}$ ES cells (Figure 5.5b,d). Rbl2 repressor was recently shown to associate with the *DNMT3B* promoter in human glioblastoma cells ([286] and Discussion). In mouse ES cells, *Rbl2* is expressed at low levels, and during neuronal differentiation its expression correlates inversely with the expression of the miR-290 cluster and *de novo* DNA methyltransferases (F.M. and D. Sch"ubeler, Friedrich Miescher Institute, unpublished results). We used RNA interference to obtain more direct evidence that *Rbl2* indeed regulates the expression of *de novo* DNA methyltransferases. Transfection of siRNAs against *Rbl2* resulted in a marked increase of *Dnmt3a2* and *Dnmt3b* expression at both mRNA and protein levels (Figure 5.5e,f). Taken together, these data argue in favor of Rbl2 as a target of the miR-290 cluster that acts as a repressor, downregulating the expression of *de novo* DNA methyltransferases.

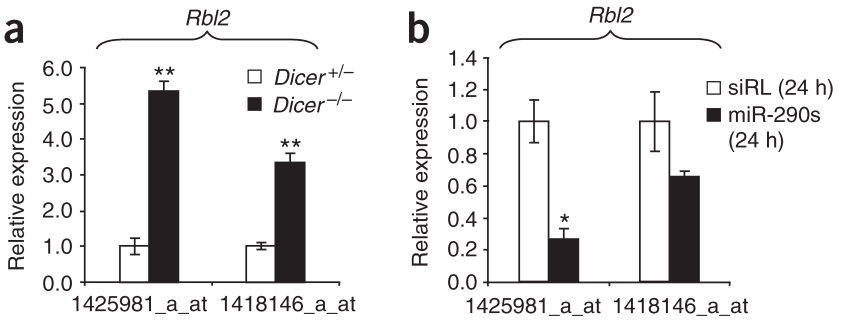


Figure 5.5: **Retinoblastoma-like protein 2 (*Rbl2*) regulates the expression of *Dnmt3a2* and *Dnmt3b*** (**a**) Levels of *Rbl2* mRNA are upregulated in *Dicer*$^{-/-}$ cells as indicated by analysis of Affymetrix arrays. The probe sets detecting expression of *Rbl2* are indicated. Mean expression values ($\pm$s.d.; n = 3) in Dicer$^{+/-}$ cells were set to one. The P-values from left to right are 0.0010 and 0.0023. (**b**) Transfection of miR-290 miRNAs into *Dicer*$^{-/-}$ ES cells downregulates the level of Rbl2 mRNA. Cells were transfected for 24 h with either a mixture of the miR-290 cluster miRNAs or with siRL (small interfering RNA against *Renilla* luciferase mRNA). Mean expression values ($\pm$s.d.; n = 3) in siRL-transfected cells were set to one. The P-values from left to right were 0.0135 and 0.1082.

5.2.4 Defective DNA methylation of *Oct4* in *Dicer*$^{-/-}$ cells

To investigate in more detail the differentiation defects in *Dicer*$^{-/-}$ cells and the possible role of the miR-290 cluster, we examined expression of *Oct4*, the core pluripotency regulator of ES cells. When differentiation was induced with 100 nM RA in the absence of leukemia inhibitory factor (LIF), the mRNA and protein levels of *Oct4* decreased similarly in *Dicer*$^{-/-}$ and control cells at day 3 (Figure 5.6a). The expression level of the orphan nuclear receptor gene *Gcnf*, an early repressor of *Oct4*, *Nanog* and

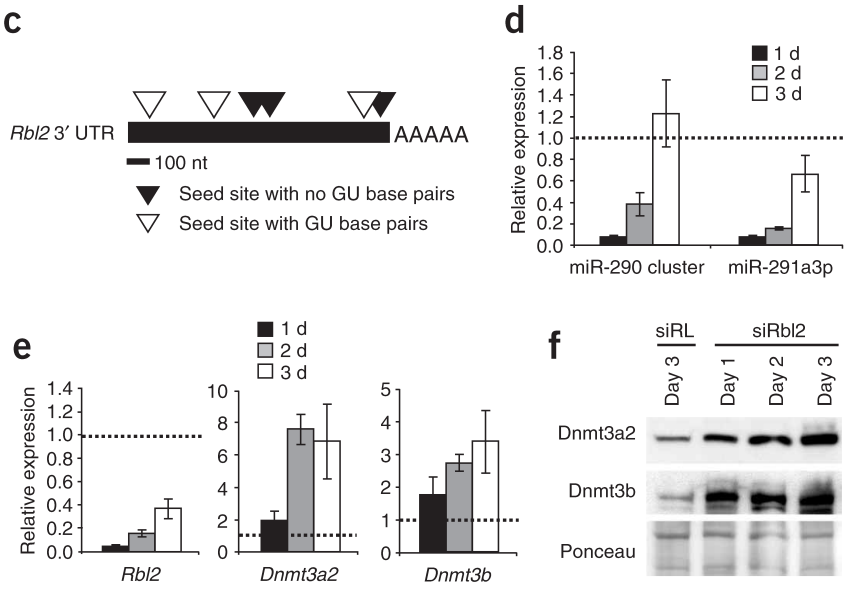


Figure 5.5: **Retinoblastoma-like protein 2 (*Rbl2*) regulates the expression of *Dnmt3a2* and *Dnmt3b*** (**c**) Schematic representation of the localization of predicted binding sites for AAGUGC seed-containing miRNAs in the 3' UTR of *Rbl2* mRNA. Predicted binding sites that contain GU base pairs in the seed and those without GU base pairs in the seed are marked with white and black triangles, respectively. (**d**) Downregulation of *Rbl2* in response to transfection of $Dicer^{-/-}$ ES cells with either all miR-290 cluster miRNAs or miR-291a-3p. For other details see Figure 5.3e. (**e**) Effect of *Rbl2* knockdown on *Dnmt3a2* and *Dnmt3b* mRNA levels. Cells were transfected with siRNAs against *Rbl2* (siRbl2) or with siRL as a control. For other details see Figure 5.3e. (**f**) Western blot analysis of Dnmt3a2 and Dnmt3b expression 1 d, 2 d or 3 d after siRbl2 transfection. Expression after transfection of siRL (3 d) is shown as a control. Ponceau staining served as a loading control.

other pluripotency markers [287], was upregulated to the same extent in *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ cells after 1 d of RA treatment (Figure 5.6b). We could also detect accumulation of repressive histone marks at the *Oct4* promoter (Figure 5.6c,d), indicating that the initiation of *Oct4* silencing was not strongly perturbed. However, repression of *Oct4* at day 6 of differentiation was clearly stronger in Dicer$^{+/-}$ ES cells (Figure 5.6e). When RA was removed at day 6 and the cells were cultured in the presence of LIF for an extra 4 d, the *Oct4* mRNA levels in *Dicer*$^{-/-}$ cells increased to approximately 40% of the initial level, whereas *Oct4* expression remained repressed in *Dicer*$^{+/-}$ cells (Figure 5.6e). A similar pattern of expression was observed for *Nanog*.

Incomplete and reversible silencing of *Oct4* in RA-treated *Dicer*$^{-/-}$ ES cells is notably similar to findings demonstrating that the stable silencing of *Oct4* is dependent on a correct *de novo methylation* of DNA [288, 289]. Therefore, we used bisulfite sequencing to analyze the methylation status of the *Oct4* promoter during the RA-induced differentiation. In *Dicer*$^{+/-}$ ES cells, DNA methylation was already detectable after 3 d of differentiation; it increased further at day 6 and remained high following the withdrawal of RA. In marked contrast, the *Oct4* promoter failed to undergo DNA methylation in differentiating *Dicer*$^{-/-}$ cells (Figure 5.6f).

To address the possibility that impaired maintenance of DNA methylation is responsible for the observed methylation defect, we analyzed several typically hypermethylated sequences and found no loss of their methylation in undifferentiated or differentiated *Dicer*$^{-/-}$ ES cells. Furthermore, expression of the maintenance DNA methyltransferase *Dnmt1* was not affected either by the loss of Dicer or upon transfection of miR-290 cluster miRNAs into *Dicer*$^{-/-}$ ES cells (Figure 5.3a-c), suggesting that maintenance of DNA methylation is not impaired in *Dicer*$^{-/-}$ ES cells.

5.2.5 Rescue of *de novo* DNA methylation of *Oct4* by miRNAs

We tested whether ectopic expression of Dnmt3a2, Dnmt3b and Dnmt3l, or transfection with miR-290 cluster miRNAs, is sufficient to rescue the defective *Oct4* promoter methylation. Co-transfection of *Dicer*$^{-/-}$ ES cells with constructs expressing all three methyltransferases from a heterologous promoter restored the *de novo* DNA methylation in *Dicer*$^{-/-}$ cells treated with RA for 3 d (Figure 5.7a). Transfection of *Dicer*$^{-/-}$ ES cells with miR-290 cluster miRNAs had a similar effect (Figure 5.7a). These results indicate that the observed *Oct4* promoter methylation defect is due to the repressed expression of *de novo* DNA methyltransferases in *Dicer*$^{-/-}$ ES cells.

To address whether the DNA methylation defect is more general, we analyzed the methylation status of two testis-specific genes, *Tsp50* and *Sox30*, which are silenced in ES cells and undergo *de novo* DNA methylation during differentiation (F.M. and D. Sch"ubeler, unpublished results). *Dicer*$^{+/-}$ but not *Dicer*$^{-/-}$ ES cells showed limited DNA methylation at *Tsp50* and *Sox30* promoters, even in the undifferentiated state.

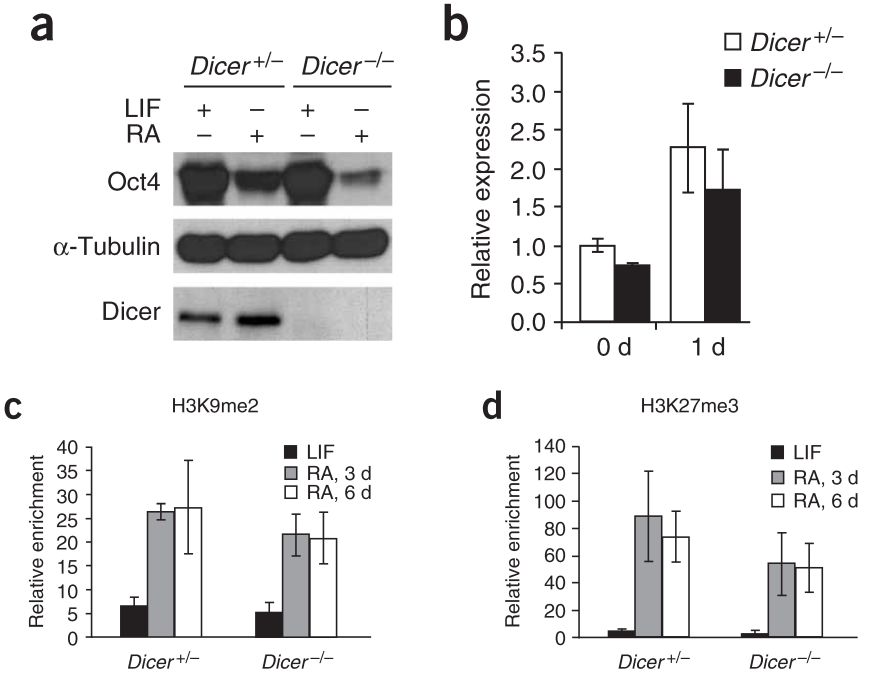


Figure 5.6: **Oct4 expression during differentiation of *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ embryonic stem (ES) cells** (**a**) Western blot analysis of Oct4 levels in *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ cells cultured in the presence of either leukemia inhibitory factor (LIF) or retinoic acid (RA) for 3 d. (**b**) Similar upregulation of the orphan nuclear receptor gene *Gcnf* expression in Dicer$^{+/-}$ and *Dicer*$^{-/-}$ cells in response to RA. Expression was estimated by RT-qPCR. The values, normalized to glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*) expression, represent means ($\pm$s.e.m.; n $\geq$ 3). Expression in control *Dicer*$^{+/-}$ cells at the 0 d time point was set as one. (**c,d**) Accumulation of repressive histone marks at the *Oct4* promoter. *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ cells, cultured in the presence of LIF, RA for 3 d (RA, 3 d) or RA for 6 d (RA, 6 d), were used for chromatin immunoprecipitation (ChIP) analysis using antibodies against dimethylated histone H3 lysine 9 (H3K9me2; **c**) and trimethylated histone H3 lysine 27 (H3K27me3; **d**). The enrichment values represent means ($\pm$s.e.m.; n $\geq$ 3).

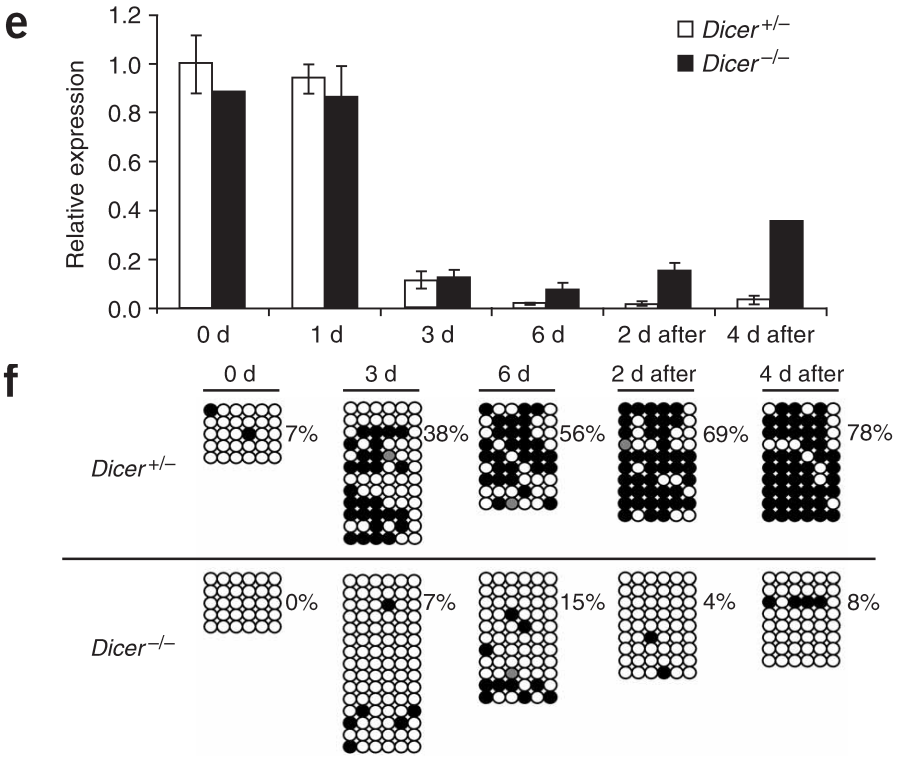


Figure 5.6: **Oct4 expression during differentiation of *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ embryonic stem (ES) cells** (**e**) RT-qPCR analysis of *Oct4* expression during RA-induced differentiation at 0 d, 1 d, 3 d or 6 d, and after returning the cells to the LIF-containing medium for up to an additional 4 days (2 d after, 4 d after). *Oct4* expression was normalized to *Gapdh* as in b ($n \geq 3$). (**f**) Analysis of CpG methylation of the *Oct4* core promoter (positions -212 to -8) during differentiation, followed by 2 d or 4 d culture in the presence of LIF. Each row of dots represents CpGs in one sequenced clone. Black dots represent methylated CpGs and white dots represent unmethylated CpGs. Sites for which the methylation status was uncertain are in gray. The cells used were the same as those used for the experiment shown in **e**. Average percentages of the methylated CpG sites are indicated.

Differentiation of *Dicer*$^{+/-}$ but not *Dicer*$^{-/-}$ cells was accompanied by additional DNA methylation. Nevertheless, the DNA-methylation changes at *Tsp50* and *Sox30* promoters were less pronounced than those observed at the *Oct4* locus, and the *de novo* DNA methylation of *Tsp50* and *Sox30* promoters was not uniformly distributed along analyzed sequences.Ectopic expression of *de novo* DNA methyltransferases affected the accumulation of DNA methylation during differentiation, whereas transfection of miR-290 cluster miRNAs resulted in increased DNA methylation at the 3' portion of the *Tsp50* sequence but had no appreciable effect at the *Sox30* promoter. Taken together, the data suggest that the defect in *de novo* methylation in *Dicer*$^{-/-}$ ES cells may be of more global character.

Dicer$^{-/-}$ ES cells grow substantially more slowly than *Dicer*$^{+/-}$ ES cells [278], and we found that transfection of miR-290 cluster miRNAs into *Dicer*$^{-/-}$ ES cells partially rescues the growth phenotype (Figure 5.7b), possibly by regulating expression of p21, an established repressor of cell-cycle progression [290]. To eliminate the possibility that the observed changes of *Oct4* DNA methylation are a consequence of different proliferation rates rather than a specific miR-290 cluster-mediated regulation, we tested whether the proliferation rate of ES cells has an effect on the onset of *de novo* DNA methylation and the expression levels of the Dnmt3 enzymes.

To reduce proliferation of *Dicer*$^{+/-}$ ES cells to a rate similar to that of *Dicer*$^{-/-}$ ES cells (Figure 5.7b), cells were treated with rapamycin, an inhibitor of mammalian target of rapamycin (TOR). Rapamycin reduces the proliferation of mouse ES cells without significantly affecting their cell-cycle profile [292], making the growth properties of rapamycin-treated Dicer$^{+/-}$ ES cells comparable to that of *Dicer*$^{-/-}$ cells [278]. In *Dicer*$^{+/-}$ cells grown in the presence of rapamycin, DNA methylation readily accumulated at the *Oct4* promoter after 3 d of RA treatment (Figure 5.7a). Likewise, decreased proliferation had no significant effect on the expression of *Dnmt3a2* or *Dnmt3b1/6*. Furthermore, restoration of *Oct4* promoter methylation by ectopic expression of *de novo* DNA methyltransferases occurred without an increase in the proliferation rate of *Dicer*$^{-/-}$ ES cells (Figure 5.7a,b). Taken together, these data demonstrate that the *Oct4* promoter methylation defect is not caused by the slower proliferation of *Dicer*$^{-/-}$ ES cells but is dependent on the miR-290 cluster miRNAs.

5.3 Discussion

Our data indicate that miRNAs bearing the AAGUGC seed, largely represented by the miR-290 cluster, are the functionally dominant miRNAs in mouse ES cells. In fact, the miR-290 cluster miRNAs were able to reverse many of the defects due to loss of Dicer when transfected into ES cells. We also found that *de novo* DNA methylation in differentiating ES cells is controlled by the miR-290 cluster and that this regulation is required for stable repression of *Oct4*. We propose that, in undifferentiated ES cells,

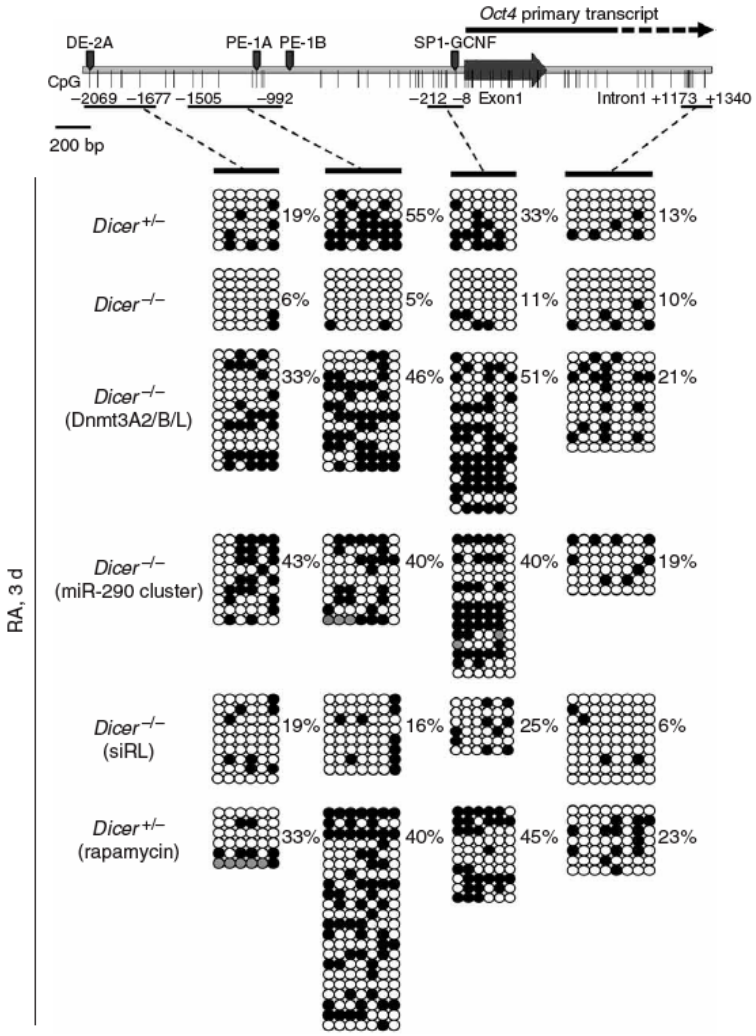


Figure 5.7: **Deficient *de novo* DNA methylation of the *Oct4* promoter in *Dicer*$^{-/-}$ embryonic stem (ES) cells can be rescued by expression of *de novo* DNA methyltransferases (Dnmts) or by transfection of miR-290 cluster miRNAs** Analysis of CpG methylation in four different *Oct4* regions. The scheme identifies positions of bisulfite-sequenced regions with respect to the *Oct4* transcription start site. SP1-GCNF depicts characterized transcription factor binding sites in the *Oct4* promoter. PE-1A and PE-1B show positions of previously characterized 1A and 1B sequences in the proximal enhancer and DE-2A is the position of 2A sequence in the distal enhancer. (for the detailed *Oct4* promoter annotation, see [291] and references therein). Represented from top to bottom: untransfected *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ cells; *Dicer*$^{-/-}$ cells co-transfected with plasmids expressing EGFP-Dnmt3a2, EGFP-Dnmt3b and EGFP-Dnmt3l; *Dicer*$^{-/-}$ cells transfected with miR-290 cluster mimics; *Dicer*$^{-/-}$ cells transfected with siRL (small interfering RNA against *Renilla* luciferase mRNA); and Dicer$^{+/-}$ cells treated with rapamycin. Both *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ ES cells were differentiated for 3 d with RA in the absence of LIF. For other details, see Figure 5.6f, the data originate from experiments independent of that shown in Figure 5.6f.

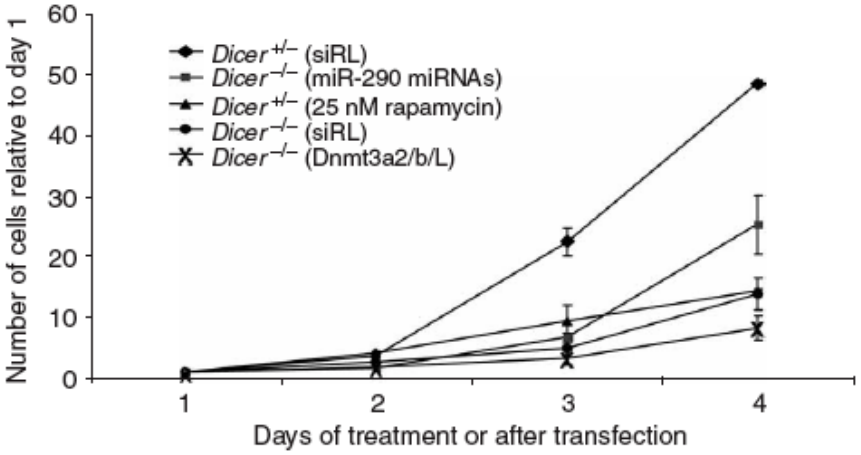


Figure 5.7: **Deficient *de novo* DNA methylation of the *Oct4* promoter in *Dicer*$^{-/-}$ embryonic stem (ES) cells can be rescued by expression of *de novo* DNA methyltransferases (Dnmts) or by transfection of miR-290 cluster miRNAs** Effects of different treatments on proliferation of *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ ES cells. Equal numbers of undifferentiated *Dicer*$^{-/-}$ and *Dicer*$^{+/-}$ cells were transfected with miR-290 cluster miRNAs, siRL or a mix of plasmids expressing Dnmt3a2, Dnmt3b and Dnmt3l. Alternatively, cells were grown in the presence of rapamycin. Average number of cells is shown relative to the number of cells present at day 1 after transfection. ($\pm$s.d.; n=3).

the miR-290 cluster miRNAs suppress a transcriptional repressor that targets genes encoding *de novo* DNA methyltransferases. The predicted primary targets of the miR-290 cluster include several transcriptional repressors, and we identified *Rbl2* as a factor contributing to repression of *Dnmt3* genes.

The expression of approximately one-quarter of predicted primary miR-290 cluster targets in ES cells is high in the oocyte but reduced in the blastocyst and somatic cells (data not shown). This resembles the situation in zebrafish, where the zygotic AAGUGC seed-containing miR-430 miRNAs control the maternal mRNA degradation [72]. However, murine maternal mRNAs are largely degraded before zygotic genome activation [280], hence before the miR-290 cluster expression. Moreover, the transition between maternal and zygotic gene expression is much slower in mammals than in the zebrafish [293]. Thus, the miR-290 cluster and related miRNAs restrict embryonic expression of genes that are highly expressed in the oocyte rather than having an extensive role in the rapid elimination of maternal transcripts. However, miR-290 cluster miRNAs and miR-430 may share some conserved roles in development, as the mouse homologs of zebrafish *lft1* and *lft2*, important regulators of mesoderm formation and targets of miR-430 ([294]), are found among $\sim$250 predicted primary targets of miR-290 cluster miRNAs.

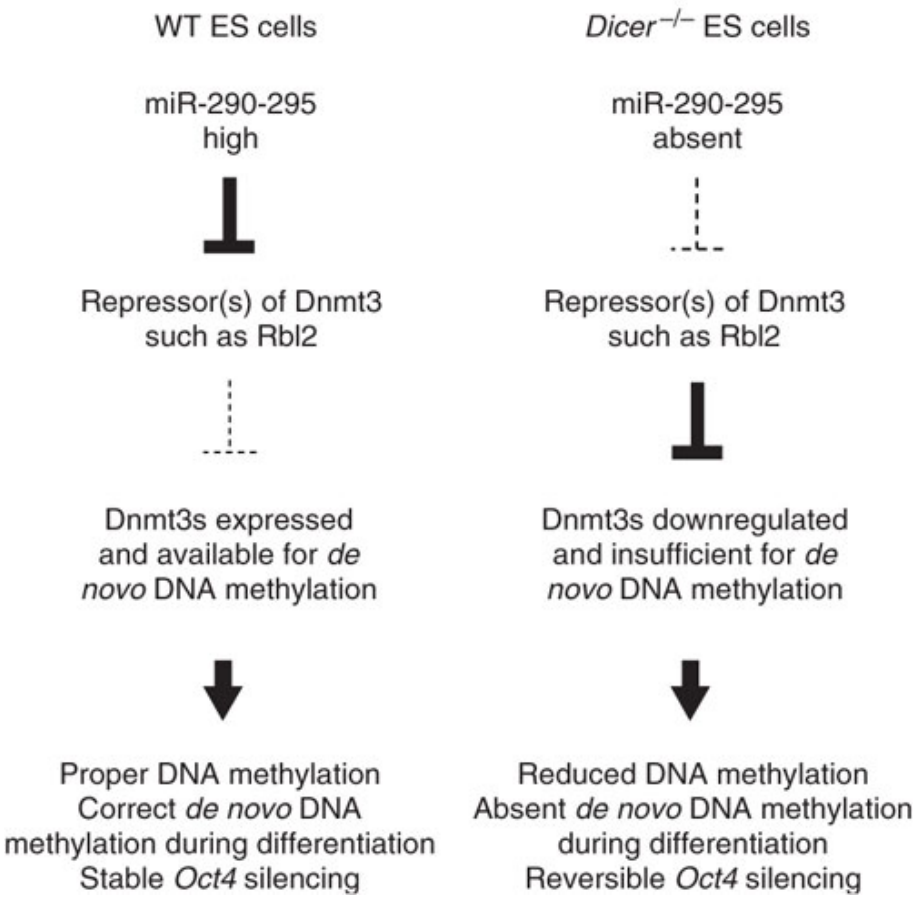


Figure 5.8: **A model for a role of miRNAs in *de novo* DNA methylation in embryonic stem (ES) cells** Dnmt, DNA methyltransferase; Rbl2, retinoblastoma-like protein 2; WT, wild type.

The microarray analysis also identified several transcripts that showed inverse changes in the Dicer knockout and miR-290 cluster rescue microarray experiments, but contained no matches to the seed of miR-290 cluster miRNAs. These are probably secondary targets whose expression is regulated by the primary targets of the miRNAs. Notably, the microarray analysis indicated that many secondary effects, probably brought about by the primary targets, are reversible despite the fact that the *Dicer*$^{-/-}$ ES cell line was established a relatively long time ago.

Both primary and secondary targets probably contribute to the reduced proliferation rate of *Dicer*$^{-/-}$ ES cells, which can be partially rescued by transfecting miR-290 cluster miRNAs. Notably, one of the predicted primary targets of the miR-290 cluster is *p21* (also known as *Cdkn1a*), a cyclin-dependent kinase inhibitor that has been shown to repress cell-cycle progression [295]. It is well established that control of *p21* expression is achieved through negative transcriptional regulators [290]. Our data argue for an additional layer of control of p21 expression by miRNAs carrying the AAGUGC seed sequence. *p21* mRNA has three GCACUU motifs in its 3' UTR, two of which are conserved across mammals. *p21* mRNA is upregulated more than three-fold in *Dicer*$^{-/-}$ ES cells, and this misregulation can be corrected by transfection of miR-290 cluster miRNAs. Thus, upregulation of *p21* could be one of the mechanisms causing the slower-growth phenotype. Although in ES cells miRNAs carrying the AAGUGC seed sequence are primarily represented by miR-290 cluster miRNAs, other related miRNAs, such as the oncomirs of the miR-17/19/106 cluster [181], could regulate expression of *p21* in other tissues. Notably, the reverse complement of AAAGUGC (positions 2-8 in miR-17-5p) was one of the motifs that was highly enriched in 3' UTRs of transcripts upregulated in human HEK293 cells depleted of Dicer or the argonaute protein AGO2 ([279]). At the same time, these cells grew more slowly, and the p21 transcript was upregulated. As miR-17/19/106 miRNAs are fairly ubiquitously expressed [170], they may provide another way to modulate expression of the *p21* tumor suppressor, with a predictable outcome for cellular growth.

The category of secondary targets includes *de novo* DNA methyltransferases, which are downregulated in *Dicer*$^{-/-}$ ES cells and upregulated upon miR-290 cluster miRNA transfection. Our data suggest that reduced expression of *Dnmt3* genes in *Dicer*$^{-/-}$ ES cells is the cause of *de novo* DNA-methylation defects observed during differentiation. Decreased expression of Dnmt3a2 and Dnmt3b, correlating with defective DNA methylation, has been described in mouse XX ES cells [296], arguing that even incomplete depletion of Dnmt3 enzymes may be limiting for proper *de novo* DNA methylation. Dnmt3a, Dnmt3b and possibly Dnmt3l may function as a complex [289]. Hence, even partial downregulation of each of them may strongly affect DNA methylation.

We investigated whether the proliferation rate itself affects *Dnmt3* expression and *de novo* DNA methylation. We found that Dnmt3 expression and *de novo* DNA methylation are not impaired when the growth of control *Dicer*$^{+/-}$ ES cells is reduced by

rapamycin. As the rapamycin-treated wild-type and *Dicer*$^{-/-}$ ES cells have comparable cell-cycle profiles and similarly slow proliferation rates [278,292], it is unlikely that the altered growth rate of *Dicer*$^{-/-}$ ES cells is responsible for decreased *Dnmt3* gene expression and the loss of *de novo* DNA methylation during differentiation. Furthermore, ectopic expression of *de novo* DNA methyltransferases rescued *de novo* DNA methylation without an apparent effect on proliferation of *Dicer*$^{-/-}$ cells. Because *de novo* DNA methylation proceeds normally inrapamycin-treated *Dicer*$^{+/-}$ ES cells, which show minimal proliferation during 3 d of RA-induced differentiation, it is unlikely that clonal effects in the cell culture would significantly distort the results of DNA-methylation analysis.

We propose that the transcription of *Dnmt3* genes is regulated in ES cells by a repressor protein whose mRNA is a target of miR-290 cluster miRNAs (Figure 5.8). Loss of the miR-290 cluster miRNAs in *Dicer*$^{-/-}$ cells would cause the upregulation of the repressor, followed by the downregulation of *de novo* DNA methyltransferases. This type of Dnmt3 regulation may be restricted to ES cells, as the levels of *Dnmt3* mRNAs are not affected in HEK293 cells with knockdown of Dicer or Argonaute proteins [279]. A suitable candidate for the repressor that targets *Dnmt3* genes is *Rbl2*, whose mRNA has all the features of a primary miR-290 cluster target. Consistent with our model, knockdown of *Rbl2* in *Dicer*$^{-/-}$ cells had a positive effect on *Dnmt3a2* and *Dnmt3b* expression. Rbl2 is a tumor suppressor that is capable of repressing E2f4 target genes as a part of the DREAM repressor complex [286]. Notably, the expression profile of human *Dnmt3b* during the cell cycle (low in G1 and G0 and upregulated in S phase) is similar to that of the E2f4 target genes repressed by Rbl2 [286]. RBL2 and the DREAM complex were recently shown to associate physically with the *Dnmt3b* promoter in human glioblastoma cells [286], suggesting that RBL2 can directly repress transcription of *Dnmt3* genes. Certainly, as the miR-290 cluster controls expression of a number of transcriptional repressors, *Rbl2* may not be the only regulator of *de novo* DNA methylation in ES cells. Fabbri et al. [297] have recently reported that the miR-29 family of miRNAs (miR-29s) can directly target *Dnmt3a* and *Dnmt3b* mRNAs and repress synthesis of *de novo* DNA methyltransferases in human lung cancer cells. miR-29 miRNAs are expressed in mouse ES cells and downregulated upon loss of Dicer, but our data argue against a major role of these miRNAs in controlling *Dnmt3a/b* mRNA or protein levels in mouse ES cells.

One of the functions of *de novo* DNA methylation during ES cell differentiation is the stable silencing of the pluripotency program. Our data indicate that, although the initial phase of transcriptional repression of *Oct4* seems to be undisturbed, the *de novo* DNA methylation of the *Oct4* promoter is severely impaired during differentiation of *Dicer*$^{-/-}$ cells. These results are consistent with the observation that stable silencing of *Oct4* is dependent on correct *de novo* methylation of DNA [288, 289]. The defect in *de novo* DNA methylation may not be confined to *Oct4*, as *Nanog*, another core

pluripotency factor, showed a similar expression profile. In addition, the promoters of *Tsp50* and *Sox30*, two testis-specific genes that are silent in ES cells and acquire de novo DNA methylation during differentiation, also failed to undergo DNA methylation in *Dicer*$^{-/-}$ cells. DNA-methylation data from these two loci are less conclusive, possibly resulting from slower kinetics of accumulation of methylation at these loci, exacerbated by a transient nature of the rescue with miR-290 cluster miRNAs. Nevertheless, accumulation of DNA methylation at these promoters is consistent with that of Oct4, suggesting a more general defect in *de novo* DNA methylation in *Dicer*$^{-/-}$ ES cells.

The defects in *de novo* DNA methylation in *Dicer*$^{-/-}$ ES cells may contribute decisively to the loss of the ability to differentiate *in vitro* and *in vivo*. Notably, *Dnmt3a*$^{-/-}$ *Dnmt3b*$^{-/-}$ double-mutant ES cells retain an undifferentiated morphology, and their late passages fail to form teratomas in nude mice [298]. The defects in *de novo* DNA methylation may also underlie the variable levels of centromeric DNA methylation reported for different *Dicer*$^{-/-}$ ES lines [277, 278], because the loss of *de novo* DNA methyltransferases results in gradual DNA demethylation during prolonged culture [298].

In summary, our analysis of gene expression in mouse *Dicer*$^{-/-}$ ES cells indicates that many of the observed transcriptome changes that occur upon loss of Dicer can be attributed to miRNAs, particularly to those of the miR-290 cluster. We have identified $\sim$250 candidate primary targets of the AAGUGC seed-containing miRNAs, and we also identified many genes that they regulate indirectly. Most notably, we demonstrated that *de novo* DNA methylation is defective in *Dicer*$^{-/-}$ ES cells, and that this is due to the indirect control of expression of the *de novo* DNA methyltransferases by the miR-290 cluster. The established link between miR-290 cluster miRNAs and *de novo* DNA methylation in ES cells indicates that miRNAs may contribute substantially to the epigenetic control of gene expression.

5.4 Methods

Cell culture.

Dicer heterozygous (+/-; line D4) and Dicer-deficient (-/-; line 27H10) ES cells (referred to as *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$, respectively) were kindly provided by G. Hannon, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA [278]. They were maintained on gelatin-coated plates with DMEM supplemented with 15% (w/v) FCS, sodium pyruvate, β-mercaptoethanol, nonessential amino acids and LIF. Differentiation of ES cells was carried out in the absence of LIF and the presence of 100 nM RA. When indicated, cells were cultured for 4 d in the presence of 25 nM rapamycin (200 μM stock of rapamycin dissolved in ethanol). Control cells were grown in the

presence of ethanol at equivalent concentration. For differentiation in the presence of rapamycin, the cells were cultured for 1 d with rapamycin and LIF followed by 3 d without LIF and with 100 nM RA and 25 nM rapamycin.

Plasmids.

The control reporter constructs encoding firefly (FL) or Renilla (RL) luciferase (pGL3-FF and pRL-TK, respectively) were described earlier [279]. FL reporters under the control of *Dnmt3b* promoter fragments (p3b-1102/+93-FF, p3b-1981/+93-FF, p3b-4997/+93-FF and p3b-7886/+93-FF) and constructs encoding the EGFP-tagged *de novo* DNA methyltransferases (pCag-EGFP-Dnmt3a2, p-Cag-EGFP-Dnmt3b and p-Cag-EGFP-Dnmt3L) were kindly provided by K. Ura, Osaka University Graduate School of Medicine, Osaka, Japan [299, 300].

Transfection of reporter constructs.

At least three independent transfection experiments in triplicate were done in each case. For luciferase assays, $Dicer^{-/-}$ cells were transfected in six-well plates with 500 ng of indicated FL reporter constructs and 50 ng of pTK-RL as a control, using Lipofectamine 2000 reagent (Invitrogen). All luciferase assays were performed 24 h after transfection.

Other transfections.

Other transfections were performed using the Mouse ES cell Nucleofection Kit (Amaxa Biosystems) and program A23 of Nucleofector I apparatus (Amaxa Biosystems). Approximately 3×10^6 $Dicer^{-/-}$ cells were used per transfection and the cells were plated immediately after electroporation. Transfections of siRNAs were performed according to the manufacturer's instructions, using 300 pmol of siRNA against RL mRNA (siRL) (Eurogentec), 50 pmol of siGENOME smartPOOL siRNAs against Rbl2 (Dharmacon), 50 pmol of each of the mmu-mir-290, mmu-mir-291a-3p, mmu-mir-292-3p, mmu-mir-293, mmu-mir-294 and mmu-mir-295 miRNA mimics (Dharmacon), or 300 pmol of mmu-mir-291a-3p, together with 2 μg of pCX-EGFP50, which served as control for transfection efficiency. For rescue of *de novo* DNA methylation by a mixture of pCag-EGFP-Dnmt3a2, pCag-EGFP-Dnmt3b and pCag-EGFP-Dnmt3L plasmids, the $Dicer^{-/-}$ cells were co-transfected with 7 μg of each of these plasmids, using the Nucleofector I apparatus. The EGFP-expressing cells were collected using a MoFlow cell sorter (Dako Cytomation) after 3 d of culture in the presence of 100 nM RA and the absence of LIF.

Chromatin immunoprecipitation.

Chromatin immunoprecipitations (ChIPs) were performed as described previously [301]. *Dicer*$^{+/-}$ and *Dicer*$^{-/-}$ ES cells, either undifferentiated or treated for 3 d with RA, were cross-linked by adding formaldehyde directly to the medium to a final concentration of 1% (w/v) at room temperature. The reaction was stopped after 8 min by adding glycine to a final concentration of 0.15 M. Cell lysates were sonicated to generate 300-1,500-bp DNA fragments. After preclearing the samples with Protein A Agarose (Upstate), the immunocomplexes were formed using anti-H3K9me2 or anti-H3K27me3 antibodies (Upstate). Immunocomplexes were collected with 30 μl of Protein A Agarose (Upstate). The purified DNA and a 1:100 dilution of the respective input DNA were used as templates for quantitative real-time PCR (RT-qPCR), using the ABI Prism 7000 Sequence Detection System (Applied Biosystems), Platinum SYBR Green qPCR SuperMix (Invitrogen) and primers specific for the glyceraldehyde-3-phosphate dehydrogenase (*Gapdh*) and *Oct4* promoters. Obtained values were first normalized to the respective input DNA and further to the enrichment at the *Gapdh* promoter where these modifications do not accumulate. Annealing of all primers was done at 55 °C.

Bisulfite sequencing.

Bisulfite sequencing was performed using the Epitect Bisulfite sequencing kit (Qiagen) according to the manufacturer's conditions. Up to 2 μg of genomic DNA was used as a starting material. PCR amplification conditions were as described [301].

Statistical analysis.

Analysis of microarray data and motifs, including statistical methods, is described in detail in the Supplementary Methods. All remaining statistical analysis used two-tailed t-tests.

Real-time quantitative RT-PCR (RT-qPCR)

Total RNA from ES cells was extracted using the Absolutely RNA Miniprep Kit (Stratagene). A Thermoscript RT-PCR kit (Invitrogen) was used for the cDNA synthesis reaction with 1 μg template RNA and 250 pmol of oligo(dT)20 primer, incubated for 1 h at 55 °C. Subsequently, cDNA was used as a template for RT-qPCR with the ABI Prism 7000 Sequence Detection System and Platinum SYBR Green qPCR SuperMix, using gene-specific primers. For Dnmt3 enzymes, splice-variant-specific primers were used. Sequences of primers are provided in Supplementary Table 4. Annealing of all primers was done at 55 °C. Relative expression levels were calculated using the formula $2^{-(\Delta Ct)}$, where ΔCt is $Ct_{(geneofinterest)} - Ct_{(GAPDH)}$ and Ct is the cycle at which the threshold is crossed. For time course experiments, the expression level at

day 0 in Dicer$^{+/-}$ ES cells or in siRL-transfected *Dicer*$^{-/-}$ cells was always set as 1 and expression levels at other time points were normalized to it.

Western blotting.

Cells were lysed in lysis buffer (30 mM Tris-HCl, pH 7.5, containing 150 mM NaCl, 1 mM MgCl2, 0.5% (v/v) Nonidet P-40, 1 mM DTT and protease inhibitors) and kept on ice for 10 min. Equal amounts of the lysed proteins were separated on polyacrylamide-SDS gels, blotted on polyvinylidene fluoride membrane and probed with the following primary antibodies: anti-Oct4 (Santa Cruz, dilution 1:2,000), anti-α-tubulin (5.2.1 Sigma, 1:10,000), anti-Dicer [D349 ([302]), 1:5,000], anti-Dnmt1 (Abcam, 1:500), anti-Dnmt3a (Imgenex, 1:250), anti-Dnmt3b (Imgenex, 1:250) and anti-RNA-polymerase II (Covance, 1:500). This was followed by incubation with secondary horseradish peroxidase-coupled antibodies. Detection was performed with ECL or ECL+ kits (Amersham).

Luciferase assays.

Luciferase assays were performed using the Dual-Luciferase Reporter Assay kit (Promega) according to the manufacturer's instructions. FL activity was normalized to RL activity expressed from pRL-TK. Normalized FL activity in cells transfected with pGL3-FF was always set as one.

Author contributions

L.S., P.S. and W.F. designed the study; L.S., P.S. and M.Z. designed the computational part; L.S. carried out most of the experiments; T.H. contributed to some experiments with *Dicer*$^{-/-}$ ES cells and most of the western analyses; C.G.A.-R. contributed to *Rbl2* knockdown and western analyses; D.G. and P.B. performed computational analyses; P.S. carried out some of the bisulfite sequencing and initial analysis of microarray data; F.M. helped with Sox30 and Tsp50 methylation analysis; L.S., P.S., M.Z. and W.F. wrote the manuscript.

Acknowledgments

We thank G. Hannon and E. Murchison (Cold Spring Harbor Laboratory, New York, USA), for providing *Dicer*$^{-/-}$ ES cells, K. Ura (Osaka University, Japan), for providing DNMT3 plasmids, A. Peters for providing antibodies, D. Schubeler for helpful suggestions and comments (both Friedrich Miescher Institute, Basel, Switzerland). We also thank E. Oakeley, H. Angliker and M. Pietrzak for their contributions to array analysis and sequencing (Friedrich Miescher Institute). P.S. is supported by the European Molecular Biology Organization (EMBO) SDIG program #1488, GAAV

IAA501110701 and the Purkynje Fellowship. P.B. is supported by the Swiss National Science Foundation (SNF) grant #3100A0-114001 to M.Z., and D.G. is supported by the Swiss Institute of Bioinformatics. L.S. is partially supported by the EC FP6 STREP program LSHG-CT-2004. The Friedrich Miescher Institute is supported by the Novartis Research Foundation.

5.5 Parts of the Supplementary Methods

5.5.1 Microarray data analysis

BioConductor Affymetrix package of the R software was used to import the CEL files from the Affymetrix Mouse Genome 430 2.0 Array. Probe set intensities were then background-corrected, adjusted for non-specific binding and quantile normalized with the GCRMA algorithm [303]. GCRMA-normalized microarray data were deposited in the GEO database (GSE7141 and GSE8503).

To extract a non-redundant set of transcripts for subsequent analyses of 3'-UTR sequences, probe sets with _s or _x tags, which map to multiple transcripts from different genes, were discarded. Then, the Affymetrix annotation from December 2006 was used to obtain the corresponding reference sequence (RefSeq [304]) for each probe set. When the Affymetrix array contained probe sets for alternative RefSeq transcripts for the same gene, we only used the RefSeq transcript with the median length 3'-UTR. Through this procedure, we obtained an n-to-1 probe set to RefSeq transcript mapping. For transcripts that had multiple probe sets, we discarded those that were deficient, as indicated by their very low variance across a set of unrelated experiments performed with different cell types using the same platform (Affymetrix Mouse Genome 430 2.0). Finally, the $\log_2$ intensities of the probe sets corresponding to a given transcript were averaged to obtain a transcript level measurement. We used Limma [305] to estimate the fold change and the corresponding p-value in the three replicate experiments for each condition.

To identify those motifs whose frequency in up-regulated (in *Dicer*$^{-/-}$) or down-regulated (in *Dicer*$^{-/-}$) ES cells transfected with miRNA mimics of the miR-290 family) 3'-UTRs is significantly different relative to the frequency in the entire set of 3'-UTRs, we extracted the set of transcripts up-regulated in the *Dicer*$^{-/-}$) cells (p-value < 0.001) and computed the relative frequency of all 7-mers in the 3'-UTRs of these transcripts compared with the entire set of 3'-UTRs represented on the microarray. For each 7-mer, we then plotted the $\log_2$(number of occurrences in up-regulated 3'-UTRs) on the x-axis, and the enrichment in up-regulated 3'-UTRs compared to the entire set of 3'-UTRs on the y-axis (Fig. 5.1b,e). We then used a Bayesian model that we previously introduced for comparing miRNA frequencies between samples [170]. Briefly, we estimate the posterior probabilities of the model that assumes that the frequency of a

given motif is different between two sets of transcripts (call this ”different” model), and the model that assumes that the frequency is the same (call this ”same” model), given the observed counts m and n of the motif among M and N total motifs in the two samples. We selected as significant those motifs that were enriched in the up-regulated or down-regulated set, respectively, with a posterior probability of the ”different” model > 0.99

Chapter 6

MicroRNA Activity Is Suppressed in Mouse Oocytes

Jun Ma[1], Matyas Flemr[2], Paula Stein[1], Philipp Berninger[3], Radek Malik[2], Mihaela Zavolan[3], Petr Svoboda[2,4], and Richard M. Schultz[1,4]

1 Department of Biology, University of Pennsylvania, Philadelphia, Pennsylvania 19104-6018, USA

2 Institute of Molecular Genetics, Academy of Sciences of the Czech Republic, Videnska 1083, 142 20 Prague 4, Czech Republic

3 Division of Bioinformatics, Biozentrum, University of Basel, and Swiss Institute of Bioinformatics, Klingelbergstrasse 50/70, 4056 Basel, Switzerland.

4 corresponding authors

published in *Current Biology, Vol 20, 265-270, 2010*

MicroRNAs (miRNAs) are small endogenous RNAs that typically imperfectly base pair with 3' untranslated regions (3' UTRs) and mediate translational repression and mRNA degradation. Dicer, which generates small RNAs in the miRNA and RNA interference (RNAi) pathways, is essential for meiotic maturation of mouse oocytes. We found that 3' UTRs of transcripts upregulated in *Dicer1*$^{-/-}$ oocytes are not enriched in miRNA binding sites, implicating a weak impact of miRNAs on the maternal transcriptome. Therefore, we tested the ability of endogenous miRNAs to mediate RNA-like cleavage or translational repression of reporter mRNAs. In contrast to somatic cells, endogenous miRNAs in oocytes poorly repressed translation of mRNA reporters, whereas their RNAi-like activity was much less affected. Reporter mRNA carrying let-7-binding sites failed to localize to P body-like structures in oocytes. Our data suggest that miRNA

function is downregulated during oocyte development, an idea supported by normal meiotic maturation of oocytes lacking *Dgcr8*, which is required for the miRNA but not the RNAi pathway (Suh *et al.* [306]). Suppressing miRNA function during oocyte growth is likely an early event in reprogramming gene expression during the transition of a differentiated oocyte into pluripotent blastomeres of the embryo.

6.1 Results and Discussion

6.1.1 Minimal Impact of MicroRNAs on Mouse Oocyte Transcriptome

The eight 5'-terminal nucleotides of a microRNA (miRNA) form a "seed," which hybridizes nearly perfectly with the target mRNA and nucleates the miRNA-mRNA interaction [182]. Whereas enrichment of motifs complementary to seeds of highly active miRNAs has been observed in 3' untranslated regions (3' UTRs) of mRNAs whose relative abundance is increased (hereafter referred to as upregulated) upon depletion of *Dicer1* [72,279,307], transcriptome analysis of *Dicer1*$^{-/-}$ metaphase II (MII) eggs did not identify any miRNA-related motifs [308]. Because transcriptome remodeling during meiosis [309] could mask upregulation of primary miRNA targets, we performed an analysis of fully grown germinal vesicle-intact (GV) *Dicer1*$^{-/-}$oocytes. Microarray profiling revealed a comparable number of upregulated (489, $p < 0.001$) and downregulated (628, $p < 0.001$) transcripts (Figure 6.1A). The magnitude of these changes was ~ 5 times smaller when compared to other studies of *Dicer1*$^{-/-}$-depleted mammalian cells [279, 307]. In fact, the loss of *Dicer1* in the oocyte caused a transcriptome change comparable to the effect of a single miRNA in embryonic stem (ES) cells (Figure 6.1A) [307].

We searched for heptamer motifs enriched in 3' UTRs of transcripts that were upregulated in the *Dicer1*$^{-/-}$ oocytes and that could explain the mRNA expression changes. One of the four motifs most significantly enriched was complementary to the seed of miR-1195 (GAACUCA, Figure 6.1B). This motif, however, is likely not associated with miRNA function, because miR-1195 was absent in deep sequencing of small RNAs from mouse oocytes [107]. Likewise, none of the predicted miR-1195 targets in the miRBase [206] was upregulated in the *Dicer1*$^{-/-}$ oocytes. Sylamer [310], an alternative approach to analyze miRNA signals in 3' UTRs, showed that none of the high-scoring motifs and none of the top five miRNA-related heptamers match seed regions of miRNAs with a cloning frequency in oocytes $> 0.1\%$.

We also examined motifs related to abundant miRNAs in transcriptomes of *Dicer1*$^{-/-}$ oocytes and ES cells. These motifs, which were selected based on deep sequencing data [107, 111], represent binding sites for more than half of all miRNAs cloned from

these cells. Interestingly, none of the motifs (including those for the let-7 family, which represents $sim30\%$ of maternal miRNAs [107,276]) showed any enrichment or any statistical bias in 3' UTRs of transcripts upregulated in *Dicer1*$^{-/-}$ oocytes. This contrasts with *Dicer1*$^{-/-}$ ES cells, where the most significant motifs match a family of highly abundant miRNAs ($\sim$25% of cloned miRNAs [111]), and several motifs corresponding to other abundant miRNAs also showed enrichment and deviation from the statistical background (Figure 6.1C).

Our data suggest limited miRNA-associated mRNA degradation in the oocyte and do not support the notion that miRNAs extensively modulate gene expression in oocytes [276, 311]. Our analysis of 3' UTRs of transcripts upregulated in *Dicer1*$^{-/-}$ oocytes does not provide evidence that the upregulation is associated with miRNA function via seed-mediated interaction with 3' UTRs. Likewise, we observed no significant enrichment of miRNA-associated motifs in 3' UTRs of intrinsically unstable mRNAs [312] and mRNAs degraded during meiosis [107]. Although miRNA binding sites were associated with specific transcript isoforms during meiotic mRNA degradation [313], it is unclear whether this observation reflects miRNA effects. It is possible that none of the maternal miRNAs is functionally dominant, and therefore none generates a strong signal, but this does not explain the low number of upregulated transcripts in *Dicer1*$^{-/-}$ oocytes. Alternatively, miRNA-mediated mRNA degradation is not robust, and the transcriptome change reflects the loss of endogenous small interfering RNAs (endo-siRNAs). We found that 42 of 489 upregulated but only 6 of the 628 downregulated transcripts in *Dicer1*$^{-/-}$ oocytes perfectly base pair with endo-siRNAs [108]. Because siRNA-guided cleavage by small RNAs requires less than complete base pairing and can occur without a perfect seed complementarity [314], it is plausible that inhibition of the RNAi pathway is the major cause of transcriptome changes in *Dicer1*$^{-/-}$ oocytes.

The idea that low activity of miRNA-mediated mRNA degradation is responsible for the absence of a miRNA signature in *Dicer1*$^{-/-}$ oocytes is supported by Suh *et al.* [306], who analyzed the maternal loss of *Dgcr8*, a component of the microprocessor complex involved in miRNA biogenesis. *Dgcr8*$^{-/-}$ oocytes show the same depletion of miRNAs as *Dicer1*$^{-/-}$ oocytes, yet the transcriptome of *Dgcr8*$^{-/-}$ oocytes is more similar to the wild-type, and mice with *Dgcr8*$^{-/-}$ oocytes are fertile, showing no meiotic spindle defects reported for *Dicer1*$^{-/-}$ and *Ago2*$^{-/-}$ oocytes. Therefore, the sterile phenotype of *Dicer1*$^{-/-}$ oocytes [276, 308] is likely due to misregulation of genes controlled by endo-siRNAs [107].

6.1.2 Endogenous miRNAs Poorly Repress Cognate mRNAs

To understand the function of maternal miRNAs, we used three sets of reporter mRNAs carrying binding sites for the endogenous miRNAs let-7a and miR-30c. let-7 is the most abundant miRNA family in the oocyte ($\sim$ 30% of maternal miRNAs

[107, 108, 276]). The miR-30 family is less abundant; it represents ∼8% of maternal miRNAs, as suggested by reverse transcriptase-polymerase chain reaction (RT-PCR) [276]. The deep-sequencing data suggest a lower abundance (∼w2.4% [107]), but such estimates are prone to errors [315]. To assess let-7 activity during oocyte growth and meiotic maturation, we used firefly luciferase reporters (Figure 6.2A) carrying a lin-41 fragment with two natural bulged let-7 binding sites (FL-2xlet-7), which were mutated in the control (FL-control) [316]. Because fully grown GV oocytes and MII eggs are transcriptionally quiescent, we microinjected in vitro-synthesized mRNAs instead of plasmid reporters. First, we compared let-7-mediated repression of FL-2xlet-7 mRNA microinjected into meiotically incompetent oocytes with repression of the FL-2xlet-7 plasmid or synthetic FL-2xlet-7 mRNA transfected into NIH 3T3 cells. FL-2xlet-7 expression was reduced by ∼40% relative to FL-control in oocytes (Figure 6.2B). Although this was less than repression of FL-2xlet-7 reporters in NIH 3T3 cells (∼50%, Figure 6.2B), it showed that reporter mRNA is repressed by endogenous let-7 in small, growing oocytes.

When FL-2xlet-7 mRNA was microinjected into fully grown GV oocytes, we observed inefficient let-7 repression, which was also found upon meiotic maturation (Figure 6.2C). This was unlikely due to insufficient amounts of endogenous let-7 miRNA because delivering the FL-2xlet-7 mRNA with a 50 molar excess of let-7a miRNA did not, in contrast to NIH 3T3 cells, improve reporter repression. Likewise, a 50 molar excess of let-7a antagomir did not increase FL-2xlet-7 expression in oocytes but did in NIH 3T3 cells.

To explore further let-7 function in oocytes, we obtained another set of reporters (Figure 6.3A), which contained three bulged let-7 sites (RL-3xB let-7) or a single perfectly complementary let-7 site (RL-1xP let-7) downstream of the *Renilla* luciferase coding sequence [317]. These two reporters are repressed to the same extent in different cell lines, but by different mechanisms [279]. The RL-1xP let-7 is cleaved by AGO2 loaded with let-7 in the middle of the duplex. The bulged sites of RL-3xB let-7 mediate translational repression and subsequent mRNA degradation. To extend the analysis to other miRNAs, we produced a similar set of reporters for miR-30c (Figure 6.3A).

Our results showed that repression of all miRNA-targeted reporters was reduced during oocyte growth (Figures 6.3B–D) despite a 3- and 5-fold increase in the amount of miR-30 and let-7, respectively, during oocyte growth [276]. This repression was presumably miRNA mediated because reporters harboring mutated miRNA binding sites (RL-3xM let-7 and RL-4xM miR-30) were not repressed (Figures 6.3B–D). Repression of perfectly complementary reporters was always significantly greater than that of their bulged versions, contrasting with data from cell lines where bulged reporters were repressed either more or equally as well [279]. This finding suggests that RNAi-like cleavage by miRNAs loaded on the AGO2-RISC complex is less affected during

oocyte growth than translational repression, which is typical for most natural mammalian miRNA targets. Target site accessibility probably partially influences reduced repression of all reporters; our data show that siRNAs target 3'UTR sequences less efficiently in the oocyte when compared to somatic cells or siRNAs targeting the coding sequence.

The miR-30 reporter was consistently better repressed than the let-7 reporter. This finding was unexpected because let-7 family constitutes ∼30% of maternal miRNAs, whereas miR-30 mRNAs are several times less abundant [107, 276]. An additional miR-30 binding site in the bulged miR-30 reporter could explain its better repression relative to the bulged let-7 reporter. However, this cannot explain differences between RL-1xP let-7 and RL-1xP miR-30 reporters. This difference may stem from secondary structures of miRNA binding sites or may reflect yet-unknown let-7-specific regulation.

Repression of the RL-4xB miR-30 reporter could involve miRNA-mediated translational repression, miRNA-mediated mRNA degradation, or a combination of both. Thus, we microinjected fully grown GV oocytes with the RL-4xB miR-30 reporter and assayed for luciferase activity and the relative amount of Luc mRNA (Figures 6.4A,B). Whereas RL-1xP miR-30 mRNA was reduced at protein and mRNA levels as expected, RL-4xB miR-30 luciferase activity was reduced ∼50%, whereas there was negligible reduction in the amount of *Luc* mRNA. This observation suggests that the remaining miRNA-mediated translational repression is uncoupled from mRNA degradation in fully grown GV oocytes. Therefore, we tested whether miRNA-targeted mRNAs localize to P bodies, cytoplasmic foci involved in miRNA-mediated mRNA degradation [316, 317]. We visualized let-7-targeted and nontargeted mRNAs via a MS2-YFP binding strategy [316]. Whereas the let-7-targeted and nontargeted reporters were uniformly distributed in the oocyte cytoplasm, only the reporter harboring functional let-7 miRNA binding sites was targeted to P bodies in NIH 3T3 cells (Figure 6.4C). This result is consistent with the loss of P bodies during oocyte growth [318].

Taken together, our data present a puzzling paradox: although mouse oocytes produce abundant RNA-induced silencing complex (RISC)-loaded miRNAs, their mRNA targets are poorly repressed. Uncoupling the loaded RISC from translational repression, however, may be an elegant solution for selective inhibition of the miRNA pathway in the oocyte because the RNAi and miRNA pathways have common components, e.g., Dicer and AGO2. Reducing miRNA activity during oocyte growth may have two roles. First, the low activity of miRNA-mediated mRNA degradation, perhaps linked to the absence of P bodies, may contribute to mRNA stability and accumulation in growing oocytes. Second, downregulation of the miRNA pathway may be required for oocyte-to-zygote transition. Abundant maternal miRNAs, such as let-7, are found in somatic cells [170]. Efficient reprogramming of somatic cells into pluripotent stem cells requires large remodeling of miRNA expression, including downregulation of "somatic" miRNAs like let-7 (reviewed in [319]). Therefore, reducing miRNA activity

may be associated with acquisition of developmental competence, and miRNAs may not be required until the zygotic genome activation is completed and the pluripotency program, which also controls miRNA expression [320], is established. From this perspective, suppression of maternal miRNA function during oocyte growth may be the first event in reprogramming the differentiated oocyte into pluripotent blastomeres of the embryo.

6.1.3 Experimental Procedures

Animals and Oocytes

Fully grown GV *Dicer1*$^{-/-}$ oocytes were obtained from 3A8 *Dicer1* conditional mice as previously described [308]. Meiotically incompetent oocytes; fully grown, GV-intact cumulus-enclosed oocytes; and MII eggs were collected, microinjected, and cultured as described [321–324]. All animal experiments were approved by the Institutional Animal Use and Care Committee and were consistent with National Institutes of Health guidelines.

mRNA Microarray Analysis

RNA was isolated from 25 fully grown GV-intact mouse oocytes and amplified as previously described [280, 325]. Oocytes for each sample were collected from an individual mouse, and four samples were generated for each group. Biotinylated complementary RNA (cRNA) was fragmented and hybridized to the Affymetrix MOE430 v2 chip, which contains ∼45,000 probe sets. All arrays yielded hybridization signals of comparable intensity and quality. Original CEL files were processed, and 3' UTR heptamer analysis was performed as described previously [279, 307].

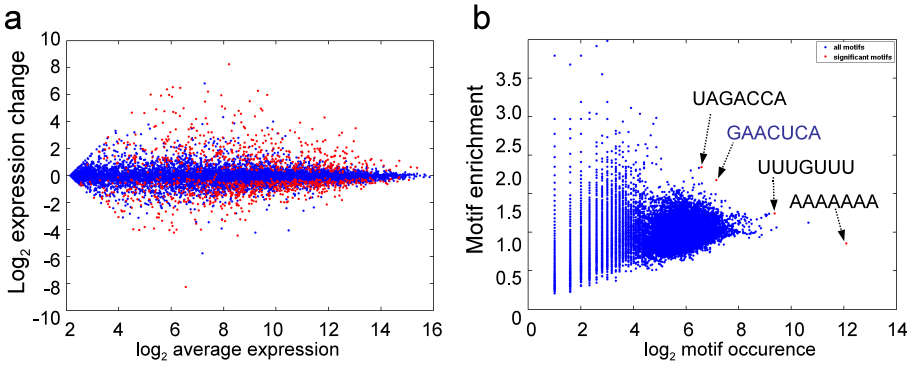
Reporter mRNA Preparation and Microinjection

Meiotically incompetent oocytes and fully grown GV oocytes were injected as described [324]. The same concentration of reporter mRNA was achieved in both stages by microinjecting incompetent oocytes with ∼1.7 pl and fully grown GV oocytes with three times that amount (i.e., ∼5 pl), because the volume of the meiotically incompetent oocytes used in these studies is about one-third of the fully grown GV oocyte. Five pl contained $\sim 10^5$ molecules of the reporter. Reporter mRNAs were microinjected at the following concentrations: FL-2xlet-7 and FL-control reporter cRNA for let-7 at 0.2 μg/μl with spiked Renilla luciferase mRNA at 0.05 μg/μl; RL-C, RL-1xP, RL-3xB, and RL-3xM for let-7 reporter at 0.05 μg/μl with spiked firefly luciferase mRNA at 0.05 μg/μl; RL-C, RL-1xP, RL-4xB, and RL-4xM for miR-30 reporter at 0.05 μg/μl with spiked firefly luciferase mRNA at 0.05 μg/ml; let-7 reporter with 12xMS2-YFP binding sites and MS2-YFP at 1 μg/μl each; let-7 mimic or antagonist at 50:1 molar ratio

to FL-2xlet-7 reporter mRNA. After microinjection, oocytes were cultured overnight in CZB containing 2.5 μM milrinone (to maintain meiotic arrest of meiotically competent oocytes) or CZB without milrinone (for meiotically incompetent oocytes) in an atmosphere of 5% CO_2 in air at 37°C before they were processed for RT-PCR analysis, luciferase assay, or immunocytochemistry.

Acknowledgments

We thank G.J. Hannon for conditional *Dicer1* knockout mice and firefly luciferase let-7 reporters, W. Filipowicz for *Renilla* luciferase let-7 reporters, and F. Duncan for help in preparing the RNA samples for microarray analysis. This research was supported by NIH grant HD22681 to R.M.S., EMBO SDIG project 1483, GACR P305/10/2215, Kontakt ME09039, and the Purkinje Fellowship to P.S.



	Dicer$^{-/-}$ oocyte	Dicer$^{-/-}$ ES cells	Dicer$^{-/-}$ ES cells +miR-290
UP	489	2551	591
DOWN	628	2578	628

Figure 6.1: **Transcriptome analysis of *Dicer1*$^{-/-}$ Oocytes.** (**A**) M [$\log_2$(fold change)] vs. A [average $\log_2$(expression level)] plot for the *Dicer1*$^{-/-}$ versus *Dicer1*$^{+/+}$ fully grown germinal vesicle oocytes. Each dot represents a transcript. Significant expression changes ($p < 0.001$ computed from four replicate experiments) are shown in red. (**B**) Heptamer motif analysis of upregulated transcripts. The motifs whose frequency in the 3' untranslated regions (3' UTRs) of upregulated transcripts is significantly different from the frequency in the entire set of 3' UTRs are shown in red. One of the significantly enriched motifs is complementary to positions 1–7 of the miR-1195.

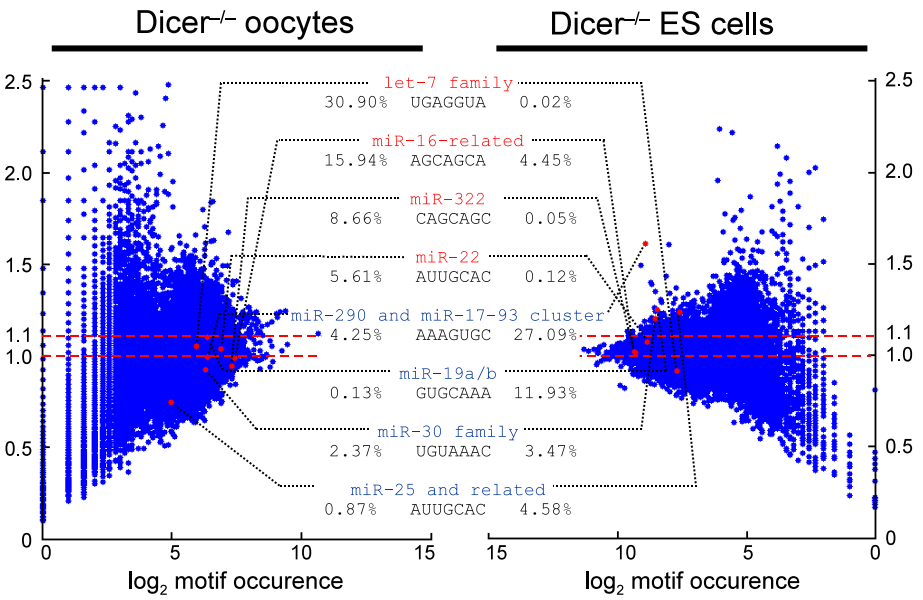


Figure 6.1: **Transcriptome analysis of *Dicer1*$^{-/-}$ Oocytes.** (**C**) Comparison of heptamer motif analyses of *Dicer1*$^{-/-}$ oocytes (left) and embryonic stem (ES) cells (right, horizontally inverted); most-relevant motifs complementary to seeds of most-abundant microRNAs (miRNAs) in both cell types are highlighted. The most-abundant miRNAs in the oocyte and ES cells are shown in red and blue text, respectively. Note that none of the motifs corresponding to abundant maternal miRNAs is enriched more than 1.1 times in 3' UTRs of transcripts upregulated in *Dicer1*$^{-/-}$ oocytes, whereas all four motifs corresponding to miRNAs abundant in ES cells are enriched in *Dicer1*$^{-/-}$ ES cells. Posterior probability analysis shows a high significance (1.000) only for the GCACUUU motif. However, posterior probability for the other three motifs corresponding to ES cell miRNAs was one to three orders of magnitude higher than all other motifs, which scored within the statistical background ($\sim 10^{-5}$). Abundance (%) of miRNAs related to individual motifs in both cell types is indicated next to each motif. Dashed lines mark 1.0- and 1.1-fold motif enrichment.

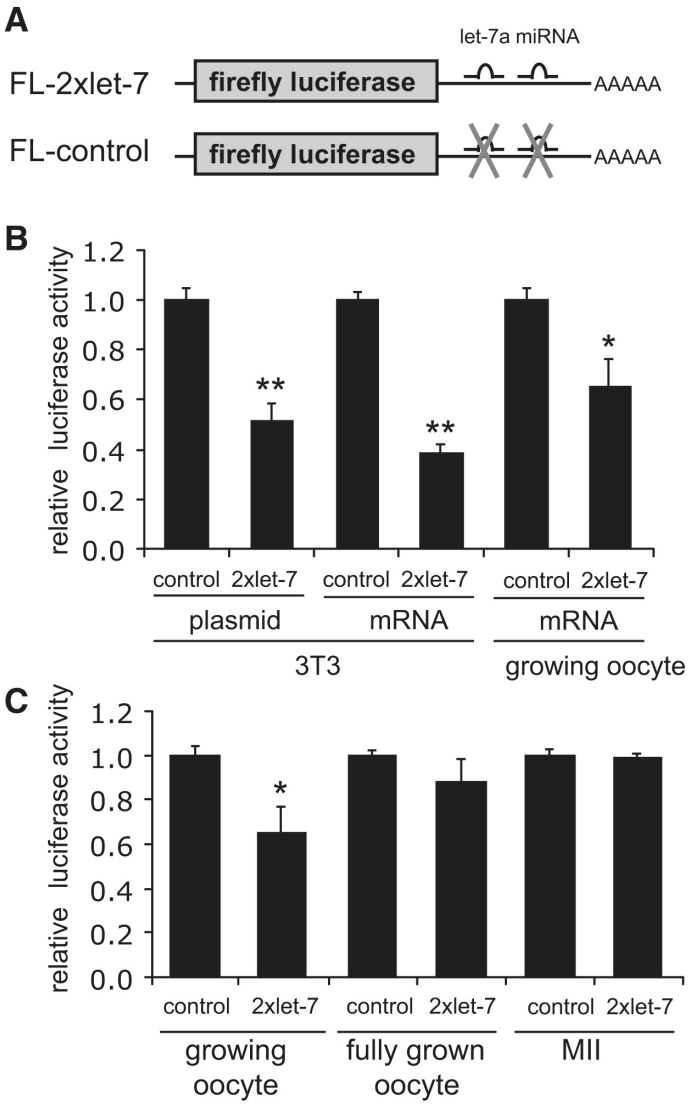


Figure 6.2: **FL-2xlet-7 Reporter Analysis.** (**A**) Schematic drawing of reporters used in experiments presented in Figure 6.2. (**B**) Relative firefly luciferase reporter activity in NIH 3T3 cells and growing oocytes. NIH 3T3 cells were transfected with reporter plasmids or mRNAs, and small, growing oocytes obtained from 13-day-old mice were microinjected with reporter mRNAs as described in the Experimental Procedures. Firefly luciferase reporter activities were normalized to the coinjected *Renilla* luciferase control and are shown relative to FL-control, which was set to one. The experiment was performed three times, and similar results were obtained in each case. Shown are data (mean $\pm$ standard error of the mean [SEM]) from one experiment. (**C**) Relative firefly luciferase reporter activity in growing oocytes obtained from 13-day-old mice, fully grown GV oocytes, and oocytes matured to metaphase II (MII). Data are presented as the mean $\pm$ SEM from six independent experiments. *$p < 0.05$, **$p < 0.01$ compared to control by analysis of variance.

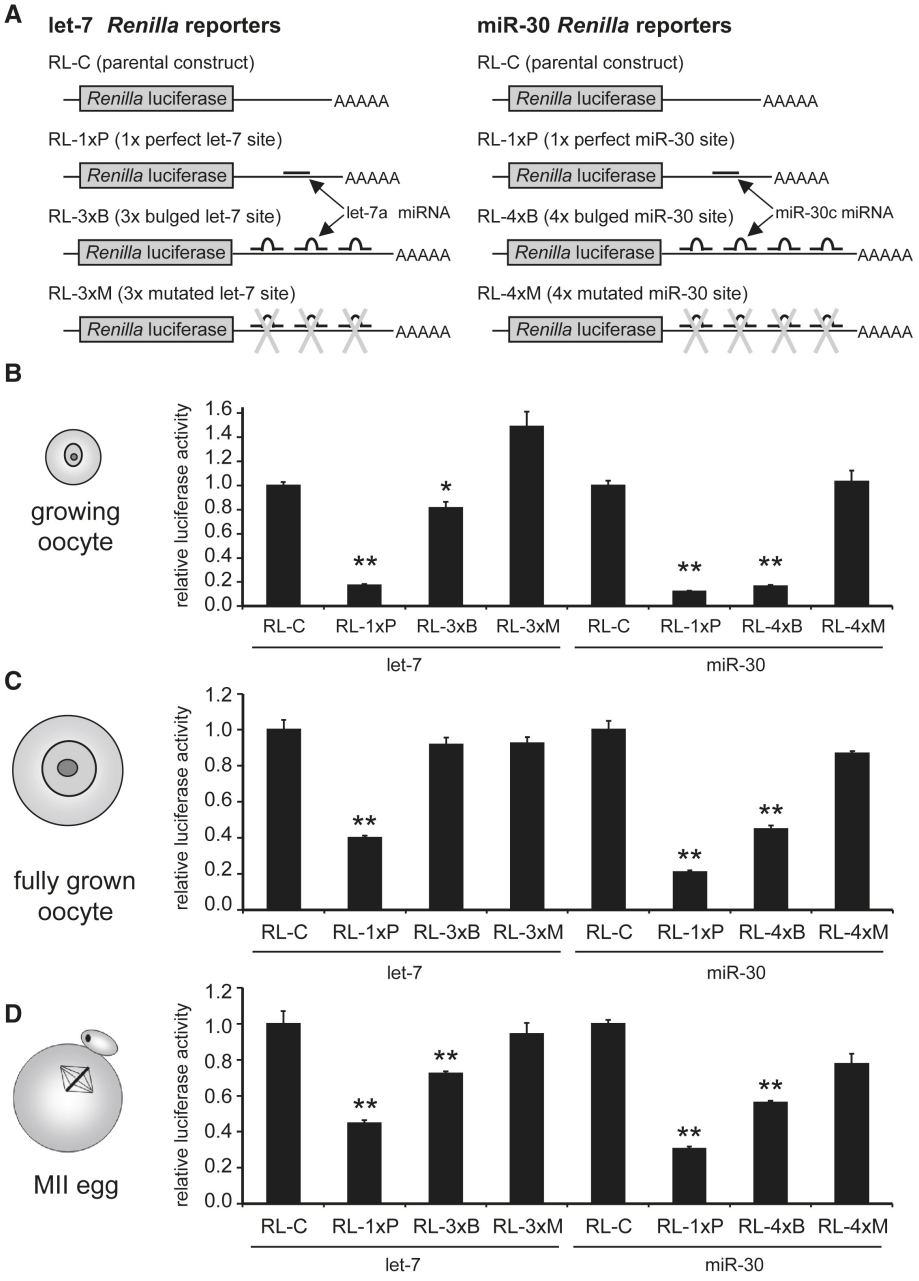


Figure 6.3: ***Renilla* Luciferase let-7 and miR-30 Reporter Analysis.** (**A**) Schematic drawing of reporters used in experiments presented in Fig. 6.3. (**B-D**). Relative *Renilla* luciferase reporter activities in growing oocytes (**B**), fully grown GV oocytes (**C**), and MII eggs (*D*). In vitro-produced reporter mRNAs were microinjected as described in the Experimental Procedures. *Renilla* luciferase reporter activities were normalized to coinjected firefly luciferase control and are shown relative to RL-C control, which was set to one for each studied miRNA. The experiment was performed three times, and similar results were obtained in each case. Shown are data (mean $\pm$6 SEM) from one experiment. *$p < 0.05$, **$p < 0.01$ compared to control by analysis of variance.

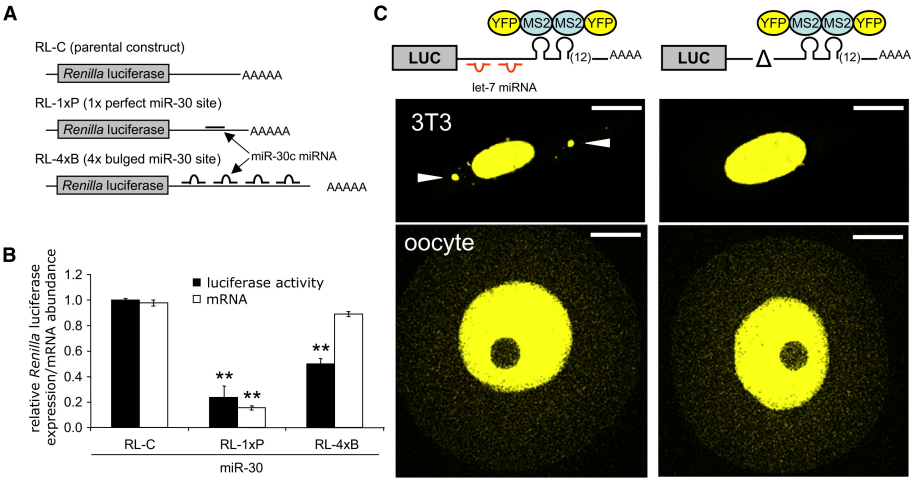


Figure 6.4: **Repressed Bulged Luciferase Transcripts Are Not Degraded and Do Not Localize to P Bodies.** (**A**) Schematic of the miR-30 reporters. (**B**). Oocytes were microinjected with miR-30 reporter mRNAs, shown in (**A**), and after 1 day of culture, the relative reporter mRNA abundance was measured by quantitative reverse transcriptase-polymerase chain reaction, and reporter mRNA translation efficiency was monitored by the dual luciferase assay. *Renilla* luciferase reporter activities were normalized to coinjected firefly luciferase control and are shown relative to RL-C control, which was set to one. The experiment was performed three times, and similar results were obtained in each case. Shown are data (mean $\pm$ SEM) from one experiment. *p < 0.05, **p < 0.01 compared to control by analysis of variance. (**C**) mRNA harboring a let-7-binding sequence fails to localize to P bodies in oocytes. Schematic depiction of reporters bound and not bound by endogenous let-7 is shown on the top of the figure. Below are confocal images showing cytoplasmic localization of corresponding reporter mRNAs. NIH 3T3 cells (top) were transfected with the corresponding reporter plasmids, and fully grown GV oocytes (bottom) were microinjected with in vitro-transcribed mRNAs as described in the Experimental Procedures. Cotransfected (coinjected) YFP-MS2 fusion protein containing nuclear localization signal is retained in the cytoplasm upon binding to reporter transcripts, thus visualizing their localization [316]. White arrowheads depict P bodies visualized by let-7-targeted reporter mRNA in NIH 3T3 cells. Scale bars represent 20 μm.

Chapter 7

Reanalysis of piRNA sequence reads reveals short byproducts of the ping-pong mechanism

Philipp Berninger and Mihaela Zavolan
Division of Bioinformatics and Swiss Institute of Bioinformatics,
Biozentrum, University of Basel, Switzerland

Research has been continued during the postdoctoral stage and published as
Conserved generation of short products at piRNA loci.
Berninger P, Jaskiewicz L, Khorshid M, Zavolan M.
BMC Genomics. 2011 Jan 19;12:46.

7.1 Introduction

During the last years, members of the Argonaute protein family have been identified as key players in small RNA guided silencing pathways. The members of the Piwi clade of Argonautes are predominantly expressed in germ cells and the association with a small RNA population distinct of miRNAs has been shown for various animals [35, 113, 114, 124, 127–129, 159]. These small RNAs, called piwi-interacting RNAs (piRNAs) which range from about 23 to 31 nt in length and show a bias for uracil at the 5' end. They are organized in clusters that represent distinct genomic loci. The infertile phenotypes observed in mouse and fruit fly carrying mutations of Piwi proteins suggested an important role of piRNAs and Piwi proteins in germ cell

development, and later their involvement in silencing retrotransposons has been revealed [134, 154]. The post-transcriptional silencing of retrotransposon transcripts is achieved by a piRNA amplification mechanism called ping-pong mechanism [36, 134]. In the ping-pong mechanism, piRNAs complementary to retrotransposons guide Piwi proteins to cleave retrotransposon transcripts. The cleavage of the target transcript occurs at the bond between the nucleotides which are opposite to the nucleotides 10 and 11 of the piRNA. This cleavage defines the 5' end of the secondary piRNA that is generated from the transposon transcript. After the 3' end is generated through nuclease cleavage and 2'-O methylation, the secondary piRNA itself is loaded into a Piwi protein and guides the cleavage of new primary piRNA precursors. This results in the production of piRNAs from the same loci from which the initial piRNAs were derived. So far, most studies on piRNAs analyzed sequence reads whose length was in the range of prototypical piRNAs (23-31 nucleotides), therefore masking out potential intermediates and/or byproducts of the amplification loop that were outside of this range. Such products could arise during the processing steps that generate both 5' and 3' ends of piRNAs and may leave detectable traces in deep sequencing libraries. This was in fact suggested by Murchison *et al.* who sequenced small RNAs of 18-30 nucleotides from platypus testes and observed that a high proportion of RNAs that were derived from piRNA loci was short relative to prototypical piRNAs. However, apart from their genomic origin those small RNAs have not been further characterized [117]. We decided to reanalyze this data set to further characterize the nature of these small RNAs and to search for processing products of the ping-pong cycle.

7.2 Results

After obtaining and mapping the small RNAs from this data set against the platypus genome, we determined genome-wide the relative distances between the 5' coordinates of perfect and uniquely mapping sequences from opposite strands (see Methods). When two sequences from opposite strands have the same 5' coordinate, this distance would be 0, whereas the primary and secondary piRNAs that are related through the ping-pong mechanism give a signature distance of 9 nucleotides. To prevent our results from being dominated by a few very abundantly sequenced clusters, we counted the pairs of sense-antisense sequences as follows. Given a sense read with copy number c_s and an antisense read with copy number c_a we count towards the distance $l_a - l_s$ the minimum of the two copy numbers. Here l_s is the genome coordinate of the sense sequence and l_a the genome coordinate of the antisense sequence. For example, if we have 3 sequences starting on the sense strand at position l_s and 7 sequences on the antisense strand at position l_a, we counted 3 observed 'pairs' at distance $l_a - l_s$.

As expected, we detected a strong signal for the ping-pong amplification cycle (see Fig. 7.1a), that is we found a high frequency of sense-antisense pairs with a distance of

9 between 5' ends. Surprisingly, we also detected a signal corresponding to a difference of 28 nt between the 5' ends, which has not been reported before (see Fig. 7.1a). Note that in this analysis we did not consider solely sequences whose length was in the range of prototypical piRNAs (23-31 nt). Because previous analyzes of piRNAs only used relatively long sequences (longer than 23 nucleotides) and did not report the peak at 28 nucleotides, we separated the sequences in different sets based on their length and repeated the analysis. Indeed, restricting the analysis to sequences longer than 23 nucleotides (the piRNA range) revealed only the peak at 9 nucleotides (P9). The peak at 28 (P28) became apparent when we analyzed sequences of disparate lengths, with one long sequence in the range of a prototypical piRNA, and a smaller ($\leq$ 23 nt) sequence on the opposite strand (see Figure 7.1b). Further analysis revealed that mostly 19 nt long sequences contributed to the short member of the pair (see Fig. 7.1a).

By inspecting some examples of pairs that gave P28 we noticed that in the P28 duplex the pairing between the 19 nt-long sequence and the long piRNA sequence started at the 11^{th} nt from the 5' end of the long sequence, as if the 3' end of the 19 nt-long sequence were defined by the ping-pong mechanism. Therefore, we decided to address systematically the question of whether the P28 and P9 co-occurred. The regions of overlap between sense and antisense pairs that gave rise to the P9 and the P28 peaks were anchored at the middle of the interval defined by the 5' ends of the sense-antisense pairs. We then performed cross-correlation analysis of those positions (see Methods). We obtained two symmetrical peaks at -9 and 10, which suggested that the middle of the P28 interval occurs preferentially either 9 nt downstream or 10 nt upstream of the middle of the P9 interval (see Fig. 7.1c,d). By separately analyzing the cases in which the short sequence of a P28-defining pair occurred on the plus and on the minus strand, we determined that the peak at -9 corresponded to cases in which the short sequence was on the plus strand and the peak at 10 to cases in which the short sequence was on the minus strand. Cross-correlation therefore showed that sense-antisense pairs with a distance of 9 between 5' ends indeed to co-occur with sense-antisense pairs with a distance of 28 between 5' ends, with an arrangement shown in Fig. 7.1e. That is, there are two long piRNA sequences from opposite strands that overlap by 10 nucleotides, characteristic for the ping-pong mechanism, and a 19 nt-long sequence which is located upstream of one of the two piRNAs (see Fig. 7.1e).

Reasoning that a meaningful processing pattern would be conserved across species, we analyzed a number of other publicly available data sets that were generated from a variety of species, such as mouse, fruit fly, zebrafish and flatworm. Unfortunately, piRNA studies in these species focused on the relatively long piRNA fraction of the total RNA. That is, only the band believed to contain sequences longer than about 20 nucleotides was sequenced. Because we expected the 19 nucleotide long reads to be much sparser in these species, we included in our analysis not only uniquely mapped sequences, but also sequences that mapped somewhat repetitively, i.e. with less than 10

loci. Of course, in this case, we weighted the copy number of a sequence coming from a given locus by the number of loci to which the sequence maps. We further focused on pairs formed by 19 nt-long sequences and prototypical piRNAs, and we did not sought evidence for the P9 peak characteristic of the ping-pong mechanism. In some of these species evidence for the existence of the ping-ping mechanism was already provided [119, 130, 134, 154]. Indeed, we found weak evidence for the P28 pattern in mouse, fruit fly, zebrafish and flatworm (see Fig. 7.2), but not in nematodes (data not shown). The nematodes are the only ones among the studied species for which no evidence for ping-pong amplification of piRNAs has been reported [159] to date. This suggests conserved mechanism for the generation of 19 nt sequences during piRNA biogenesis.

7.3 Discussion

A reanalysis of the piRNA sequence reads in five species surprisingly revealed that the signature of the ping-pong mechanism is accompanied by 19 nt long sequences upstream of piRNAs. The biogenesis of these 19-mers is still elusive. Although we cannot exclude that they emerge during the production of primary piRNAs, because their production appears to be associated with the production of what appear to be piRNA-guided cleavage fragments of Piwi proteins, we speculate that the 3' ends of these 19-mer sequences are most likely generated during the same cleavage event that produces the 5' ends of the secondary piRNAs. What defines the 5' end of these 19-mers is yet unknown. Our current hypothesis is that a 5'$\rightarrow$3' exonuclease, which stops when it reaches a double-stranded complex formed by the primary piRNA and the future 19-mer fragment or when it reaches the Piwi protein (that holds the piRNA-target duplex), is responsible for the formation of the 5' ends of these sequence.

It is unclear whether these 19-mers have a specific function or are simply by-products of piRNA biogenesis. If they are incorporated into Piwi proteins, they would allow the ping-pong mechanism to move on the transcript, instead of being limited to the location defined by the primary piRNA because the cleavage that would be guided by the 19-mer would occur at a different position in the transcript relative to the cleavage that is induced by the secondary piRNA located immediately downstream of the 19-mer. Even as degradation products, understanding how the 5' ends of these sequences are generated will shed further light into the piRNA biogenesis. Finally, because currently we cannot define precisely what a piRNA is (other than using as criteria the length of the sequence and its association with the Piwi proteins), the P28 and P9 patterns used in conjunction may allow us to more precisely identify true piRNAs in deep sequencing samples of small RNAs.

In order to follow up on these ideas, we would like to know how abundant the 19 nt sequences are in testis lysate and immunoprecipitated Piwi proteins compared with

primary and secondary piRNAs. This is because, as we already noticed, none of the data sets that are publicly available and we analyzed were generated optimally to allow us to identify the 19-mers. Towards this end we will isolate total RNA from mouse testes and we will perform deep sequencing of the small RNA fraction in the range of 15-40 nucleotides. To investigate whether the 19-mers may function as piRNAs, we will also determine whether they carry a 2'-O methylation at their 3' ends. We hope that these studies will open new avenues in the identification of factors that are responsible for piRNA biogenesis.

7.4 Methods

7.4.1 Sequence annotation

From the GEO database (http://www.ncbi.nlm.nih.gov/geo/) we obtained the following publicly available data sets that were obtained in previous studies [23, 24, 117, 119, 120, 130, 134, 141, 154, 158, 326, 327]: GSM266831 and GSM266832 for platypus, GSM405492, GSM406408, GSM405490, GSM405493 and GSM406409 for flatworm, GSM319953, GSM319955, GSM319956, GSM319957, GSM319959, GSM319960, GSM433288, GSM433290, GSM433292, GSM433294, GSM400967, GSM400968 and GSM179088 for mouse, GSM171830, GSM171831, GSM315420, GSM315421, GSM315422, GSM315423 and GSM315424 for zebrafish and GSM154618, GSM154620, GSM154621, GSM154622, GSM231091, GSM378200, GSM379050, GSM379052, GSM379054, GSM379056, GSM379058, GSM379060, GSM379061, GSM379063 and GSM379065 for fruit fly.

After downloading the data sets, the adaptors were trimmed according to [195], and afterwards all sequences of minimum length 15 nt were aligned against the corresponding genomes with oligomap [195]. For platypus, zebrafish, fruit fly and mouse, we obtained the genome assemblies ornAna1, danRer7, dm3 and mm9 from the UCSC genome website (http://genome.cse.ucsc.edu) and the genome assembly version 3.1 for flatworm was obtained from the genome center at washington university (genome.wustl.edu).

After mapping the reads to the respective genomes we only use those that mapped to less than 10 loci in the case of mouse, zebrafish, fruit fly and flatworm, and only sequences that mapped uniquely in the case of platypus.

7.4.2 Position Correlations

We wanted to assess the correlation between the locations of the 5' ends of sequences deriving from opposite strands. We measured the distance Δ between the 5' ends for a given locus as follows:

$$pairs(\Delta)^{+-} = \sum \min(weight^{+}(x_i), weight^{-}(x_i + \Delta)) \tag{7.1}$$

where $weight_a^+(x)$ is the sum over all of sequences that have their 5' end on the plus strand at a particular position x of the sequence copy number divided by the number of genomic loci corresponding to the sequence. This measurement focuses only on the distance between the 5' ends and the length of the sequences is ignored. The 10 nucleotide overlap between the 5' ends of sequences from opposite strands that are generated by the ping-pong mechanism corresponds to Δ=9. The results of this computation can be interpreted quite intuitively, it just indicates how often a particular distance has been observed in the entire set of reads. On the other hand, our measure is more conservative than that proposed by Olson *et al.* [328], who multiplied the counts of sequences overlapping from sense and antisense strands, thereby putting more weight on distances observed in genomic loci with a large number of reads.

For the systematic analysis of the relative position of the two peaks that we identified in the 'pairs' analysis, we used cross-correlation analysis (matlab fuction crosscorr). First, we identified the sequence reads that give rise to P28 and P9 patterns, respectively. For each of these patterns, we generated independently a vector in which the index was the genomic location of the nucleotide that was located midway between the 5' end of the sequence on the plus strand and the 5' end of the sequence on the minus strand, and the entries were the numbers of pairs associated with a given genomic location. We then applied cross-correlation analysis between the two vectors in a window of length 50. This resulted in two peaks, one at -9 and the other at 10. We then repeated this analysis using the same vector for P9, but two different vectors for P28. One of these vector was constructed from pairs containing the 19-mer on the plus strand and the other from pairs containing the 19-mer on the minus strand. With the first vector we obtained only the peak at -9 and with the second vector only the peak at 10.

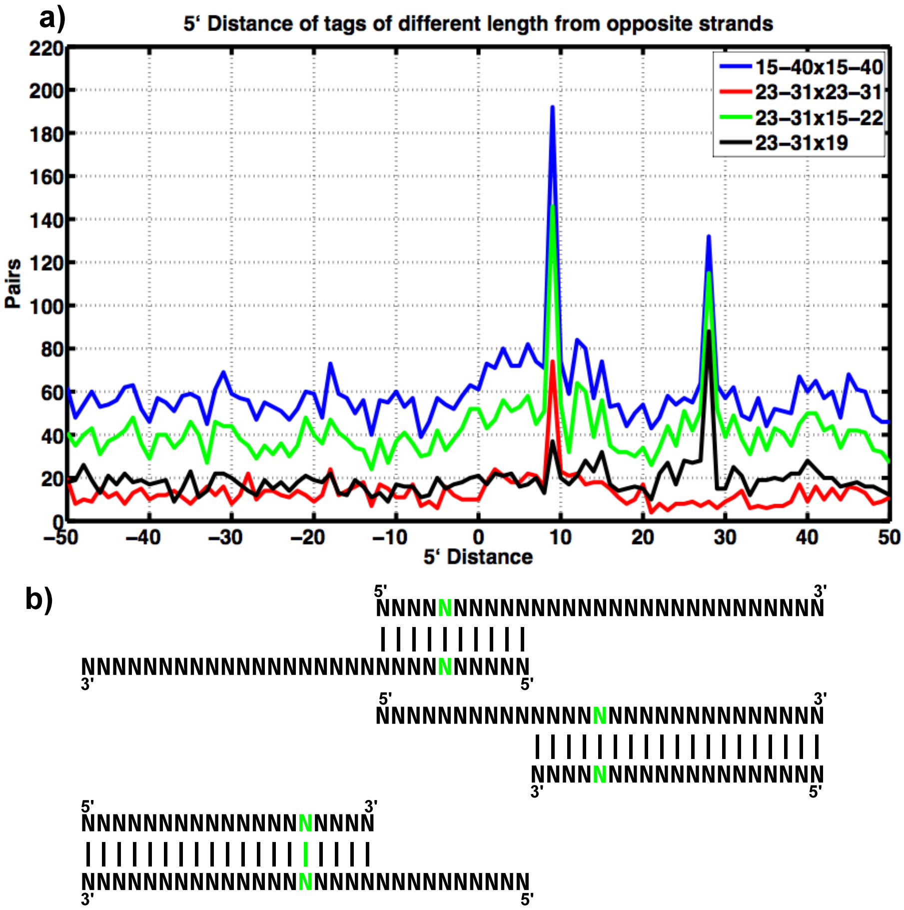


Figure 7.1: **Novel processing pattern revealed by overlapping tags.** (**a**) Overlap analysis of sequences from opposite strands inside the piRNA clusters. On the x-axis, the 5' offset of sequences deriving from opposite strands is shown, and on the y-axis, the number of detected 'pairs'. The analysis was carried out for several sets (defined by the length of reads taken into account) of perfectly and uniquely mapping sequences. For the blue line all sequences of length 15-40 nt were taken into account. The red line corresponds to sequences in the range of prototypical piRNAs (23-31 nt). For the green line, pairs were only counted, if the sequences on one strand were in the range of piRNAs (23-31 nt) and those on the opposite strand below that range (15-22 nt). In black, only pairs generated by a 19 nt long sequence on the one strand, and 23-31 nt long sequences on the other strand were taken into account. The peak at 9 (P9) is specific to a set of sequences in the piRNA range, and the peak at 28 (P28) comes from pairs in which one sequence is long and the other is short. In **b** the corresponding duplexes are shown. (**b**) On the top, the duplex corresponding to the peak at 9 is shown. Here, the opposite strand sequences overlap by 10 nt with a distance of 9 between 5' ends, which is characteristic for the ping-pong mechansim. The nucleotide located midway between the two 5' ends, which is used for the cross-correlation in **c**, is colored in green. The two duplex corresponding to the peak at 28 are shown below. The 'center' of the duplex is marked green.

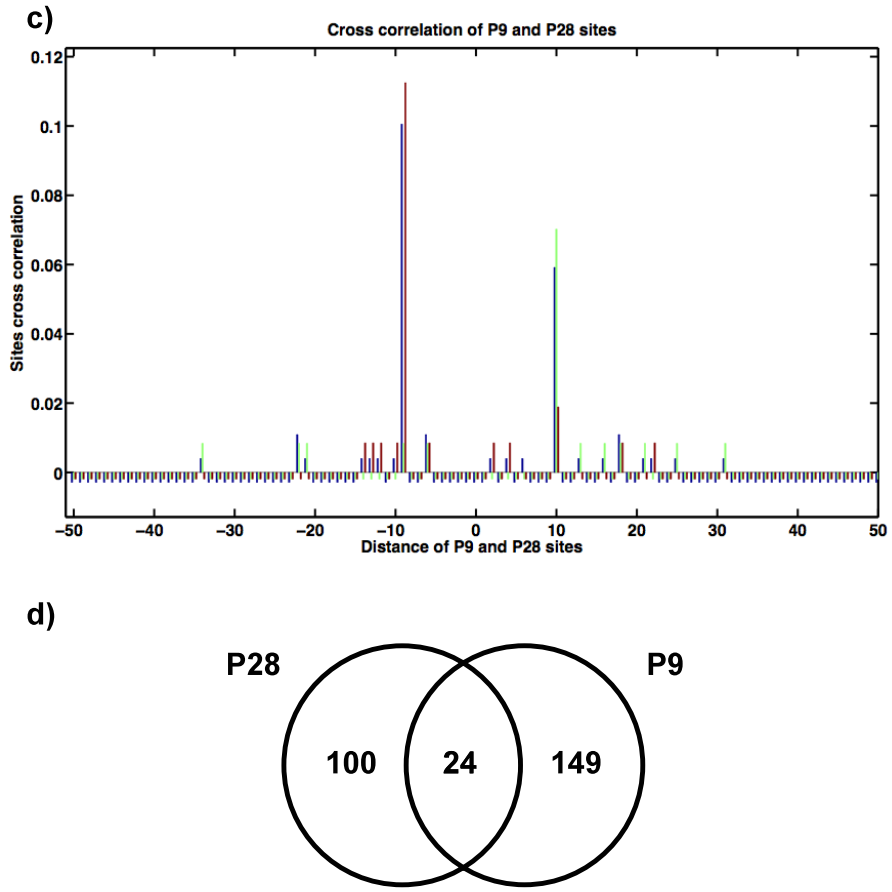


Figure 7.1: **Novel processing pattern revealed by overlapping tags.** (**c**) Cross-correlation analysis was applied on the loci responsible for the peak at 9 against the loci responsible for the peak at 28. The cross-correlation was made between the 'centers' of the regions giving rise to the P9 and P28 peaks as shown in **b**. Taking all sites together, two preferential locations of the 'centers' are found (blue). By separating the locations of the peaks of 28 based on the strand on which the 19-mer contributing to the P28 pattern occurred, the two peaks were disentangled. For brown we used P28 patterns in which the short sequence was found on the plus strand. For red we used P28 patterns in which the short sequence was found on the minus strand. (**d**) A Venn diagram showing the co-occurrence of P9 and P28 patterns. All locations on the genome that gave rise to P9 or P28 patterns were taken into account. The intersection shows that a substantial proportion of cases of P9 pattern co-occur with the P28 pattern.

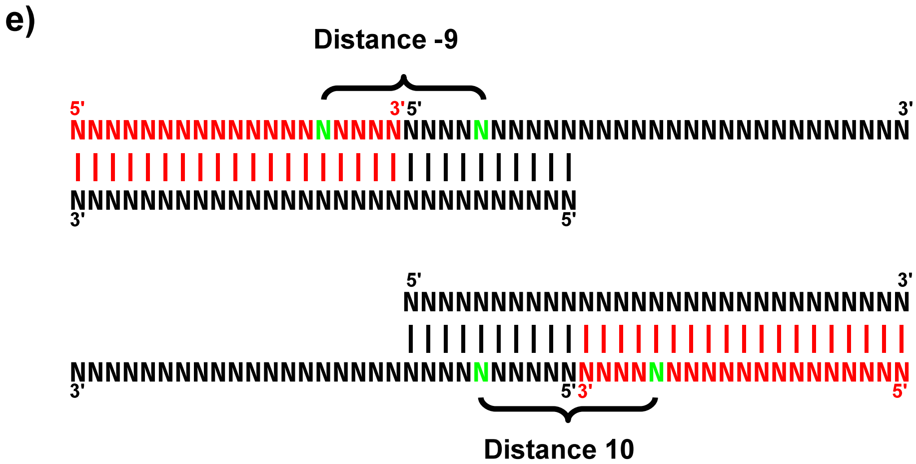


Figure 7.1: **Novel processing pattern revealed by overlapping tags.** (**e**) The duplexes, which occur if the P28 and P9 patterns co-occur are shown. The 19 nt sequence is colored in red. The top duplex corresponds to the distance of -9 of the two 'centers', the bottom one to the distance of 10. Note, the ping-pong pairs stay at the same location, only the 19 nt sequence changes place in a point symmetrical manner.

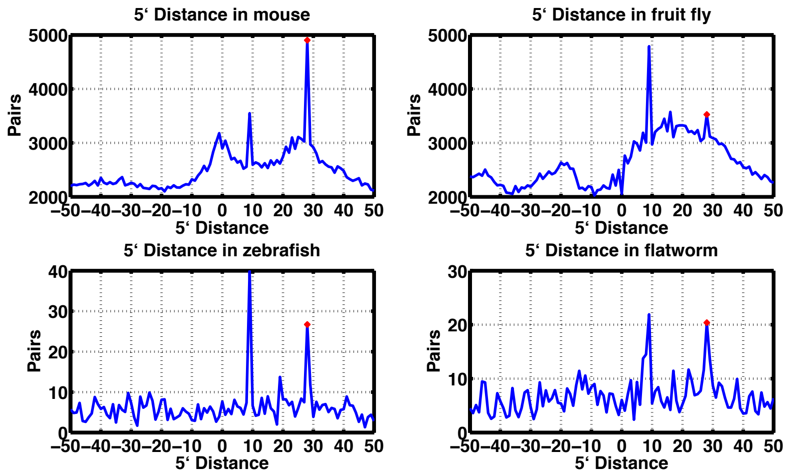


Figure 7.2: **Occurrence of the novel processing pattern is conserved** The presence of the novel processing pattern is shown for mouse, fruit fly, zebrafish and flatworm. The analysis was restricted to pairs formed by 23-31 nt sequences on the one, and 19 nt sequences on the other strand. Only sequences mapping to less than 10 genomic loci were taken into account. The position of the P28 pattern is marked with red.

Chapter 8

Conclusions

Presented in the chapters before was the work carried out to characterize small regulatory RNAs as well as their processing products and targets from large-scale datasets. We have developed an annotation pipeline, which allows both the annotation of small RNAs obtained from deep-sequencing runs, as well as the identification of small RNAs whose expression changes between different conditions or cell types. The initial step of this analysis is the identification of the genomic loci from which the sequence reads emerge and their careful classification into different functional RNA categories, which is crucial for further interpretations of the obtained data. We have connected our annotation pipeline with the Elmmo miRNA target predictions and made them available as a web server for the scientific community. This provides the user with statistical analysis and data mining tools on miRNA expression profiles and predicted miRNA target sites in several species. This in turn allows the user to focus on promising miRNA targets. Interestingly, the statistical tools which are used for the expression profile comparisons turned out to be applicable for the motif enrichment analyzes in the PURE-CLIP and $Dicer^{-/-}$ studies.

Although the additional information which is gained by the intersection between predicted miRNA targets and miRNA expression profiles results is a big improvement over target predictions alone, the experimental validation of miRNA targets still remains. In collaboration with the group of Thomas Tuschl at Rockefeller University, we have developed a method which allows identification of miRNA target sites directly from deep-sequencing data. This approach, which employs crosslinking of mRNAs and immunoprecipitation of Argonaute protein complexes, makes use of photoreactive 4-thiouridines, which are incorporated into nascent transcripts. To our surprise, we not only detected a strong motif enrichment which corresponded to the high abundant miRNAs, but also that those motifs occured predominantly 1-2 nt downstream of the crosslink site, suggesting that the crosslink occurs precisely opposite to the 9^{th} and 10^{th} position of the miRNA.

In collaboration with the groups of Witold Filipowicz and Petr Svoboda we explored the functional consequences of loss of *Dicer* on early development. The transition from oocyte to embryo is accompanied by a extensive transcriptome remodelling. Maternally inherited transcripts are degraded after fertilization and new transcripts are generated at the onset of transcription. *Dicer*$^{-/-}$ in murine embryonic stem cells revealed an important role for miRNAs. Our analysis suggested that the members of the miR-290 cluster have a major impact on the transcriptome, and through combination of experiments and computational transcriptome analysis we were able to identify primary and secondary targets of the miR-290 cluster.

In clear contrast, the loss of *Dicer* in oocytes did not point towards a disruption of the miRNA pathway. Consistent with this, we did not detect a signal for miRNA in the microarray analysis. Instead, we identified $\sim$ 10 % of the up-regulated genes as targets of endo-siRNAs. This suggests that miRNAs do not extensively modulate the gene expression in oocytes and disruption of the endo-siRNA pathway is responsible for the misregulation of genes observed in *Dicer*$^{-/-}$ oocytes.

Finally, the analysis of piRNA reads revealed the presence of 19-mer sequences upstream of ping-pong amplified piRNAs. This pattern, which turned out to be conserved among species with ping-pong amplification, might be useful for the understanding of piRNA biogenesis. It does not only provide a starting point for biochemical experiments but also offers a criterion together with the already known piRNA pattern to identify ping-pong events in deep sequencing samples of small RNAs. Processing analysis similar to ours identified tasiRNAs and methylation patterns in plants as well as the ping-pong mechanism in *Drosophila melanogaster*. Hence, it is tempting to speculate that processing analyses in the age of deep-sequencing data might reveal more valuable biological insights.

Bibliography

[1] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811, 1998.

[2] 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(5):843–854, 1993.

[3] B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–862, 1993.

[4] 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(6772):901–906, 2000.

[5] A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–89, 2000.

[6] R. C. Lee and V. Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–864, 2001.

[7] 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(5543):858–862, 2001.

[8] M. Lagos Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–858, 2001.

[9] A. J. Hamilton and D. C. Baulcombe. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–952, 1999.

[10] S. M. Elbashir, W. Lendeckel, and T. Tuschl. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev*, 15(2):188–200, 2001.

[11] A. Grishok, A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun, and C. C. Mello. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106(1):23–34, 2001.

[12] G. Hutvágner, J. McLachlan, A. E. Pasquinelli, E. Bálint, T. Tuschl, and P. D. Zamore. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–838, 2001.

[13] R. F. Ketting, S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. Plasterk. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev*, 15(20):2654–2659, 2001.

[14] H. Lin and A. C. Spradling. A novel group of pumilio mutations affects the asymmetric division of germline stem cells in the Drosophila ovary. *Development*, 124(12):2463–2476, 1997.

[15] K. Bohmert, I. Camus, C. Bellini, D. Bouchez, M. Caboche, and C. Benning. AGO1 defines a novel locus of Arabidopsis controlling leaf development. *EMBO J*, 17(1):170–180, 1998.

[16] B. Moussian, H. Schoof, A. Haecker, G. Jürgens, and T. Laux. Role of the ZWILLE gene in the regulation of central shoot meristem cell fate during Arabidopsis embryogenesis. *EMBO J*, 17(6):1799–1809, 1998.

[17] Erbay Yigit, Pedro J Batista, Yanxia Bei, Ka Ming Pang, Chun-Chieh G Chen, Niraj H Tolia, Leemor Joshua Tor, Shohei Mitani, Martin J Simard, and Craig C Mello. Analysis of the C. elegans Argonaute family reveals that distinct Argonautes act sequentially during RNAi. *Cell*, 127(4):747–757, 2006.

[18] L. Joshua Tor. The Argonautes. *Cold Spring Harb Symp Quant Biol*, 71:67–72, 2006.

[19] Marcin Nowotny and Wei Yang. Structural and functional modules in RNA interference. *Curr Opin Struct Biol*, 19(3):286–293, 2009.

[20] Julia Höck and Gunter Meister. The Argonaute protein family. *Genome Biol*, 9(2):210, 2008.

[21] Brent Brower Toland, Seth D Findley, Ling Jiang, Li Liu, Hang Yin, Monica Dus, Pei Zhou, Sarah C R Elgin, and Haifan Lin. Drosophila PIWI associates with chromatin and interacts directly with HP1a. *Genes Dev*, 21(18):2300–2311, 2007.

[22] Jianquan Wang, Jonathan P Saxe, Takashi Tanaka, Shinichiro Chuma, and Haifan Lin. Mili interacts with tudor domain-containing protein 1 in regulating spermatogenesis. *Curr Biol*, 19(8):640–644, 2009.

[23] Michael Reuter, Shinichiro Chuma, Takashi Tanaka, Thomas Franz, Alexander Stark, and Ramesh S Pillai. Loss of the Mili-interacting Tudor domain-containing protein-1 activates transposons and alters the Mili-associated small RNA profile. *Nat Struct Mol Biol*, 16(6):639–646, 2009.

[24] Vasily V Vagin, James Wohlschlegel, Jun Qu, Zophonias Jonsson, Xinhua Huang, Shinichiro Chuma, Angelique Girard, Ravi Sachidanandam, Gregory J Hannon, and Alexei A Aravin. Proteomic analysis of murine Piwi proteins reveals a role for arginine methylation in specifying interaction with Tudor family members. *Genes Dev*, 23(15):1749–1762, 2009.

[25] Chen Chen, Jing Jin, D. Andrew James, Melanie A Adams Cioaba, Jin Gyoon Park, Yahong Guo, Enrico Tenaglia, Chao Xu, Gerald Gish, Jinrong Min, et al. Mouse Piwi interactome identifies binding mechanism of Tdrkh Tudor domain to arginine methylated Miwi. *Proc Natl Acad Sci U S A*, 2009.

[26] Yohei Kirino, Namwoo Kim, Mariàngels de Planell Saguer, Eugene Khandros, Stephanie Chiorean, Peter S Klein, Isidore Rigoutsos, Thomas A Jongens, and Zissimos Mourelatos. Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability. *Nat Cell Biol*, 11(5):652–658, 2009.

[27] Jin-Biao Ma, Yu-Ren Yuan, Gunter Meister, Yi Pei, Thomas Tuschl, and Dinshaw J Patel. Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature*, 434(7033):666–670, 2005.

[28] James S Parker, S. Mark Roe, and David Barford. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434(7033):663–666, 2005.

[29] Yanli Wang, Stefan Juranek, Haitao Li, Gang Sheng, Thomas Tuschl, and Dinshaw J Patel. Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature*, 456(7224):921–926, 2008.

[30] Benjamin P Lewis, I hung Shih, Matthew W Jones Rhoades, David P Bartel, and Christopher B Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, 2003.

[31] Alexander Stark, Julius Brennecke, Robert B Russell, and Stephen M Cohen. Identification of Drosophila MicroRNA targets. *PLoS Biol*, 1(3):E60, 2003.

[32] Susanne Till, Erwan Lejeune, Rolf Thermann, Miriam Bortfeld, Michael Hothorn, Daniel Enderle, Constanze Heinrich, Matthias W Hentze, and Andreas G Ladurner. A conserved motif in Argonaute-interacting proteins mediates functional interactions through the Argonaute PIWI domain. *Nat Struct Mol Biol*, 14(10):897–903, 2007.

[33] Keita Miyoshi, Hiroko Tsukumo, Tomoko Nagami, Haruhiko Siomi, and Mikiko C Siomi. Slicer function of Drosophila Argonautes and its involvement in RISC formation. *Genes Dev*, 19(23):2837–2848, 2005.

[34] Tim A Rand, Sean Petersen, Fenghe Du, and Xiaodong Wang. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell*, 123(4):621–629, 2005.

[35] Kuniaki Saito, Kazumichi M Nishida, Tomoko Mori, Yoshinori Kawamura, Keita Miyoshi, Tomoko Nagami, Haruhiko Siomi, and Mikiko C Siomi. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. *Genes Dev*, 20(16):2214–2222, 2006.

[36] Lalith S Gunawardane, Kuniaki Saito, Kazumichi M Nishida, Keita Miyoshi, Yoshinori Kawamura, Tomoko Nagami, Haruhiko Siomi, and Mikiko C Siomi. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science*, 315(5818):1587–1590, 2007.

[37] Jidong Liu, Michelle A Carmell, Fabiola V Rivas, Carolyn G Marsden, J. Michael Thomson, Ji-Joon Song, Scott M Hammond, Leemor Joshua Tor, and Gregory J Hannon. Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, 305(5689):1437–1441, 2004.

[38] Gunter Meister, Markus Landthaler, Agnieszka Patkaniowska, Yair Dorsett, Grace Teng, and Thomas Tuschl. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell*, 15(2):185–197, 2004.

[39] Nasser Tahbaz, Fabrice A Kolb, Haidi Zhang, Katarzyna Jaronczyk, Witold Filipowicz, and Tom C Hobman. Characterization of the interactions between mammalian PAZ PIWI domain proteins and Dicer. *EMBO Rep*, 5(2):189–194, 2004.

[40] Sam Griffiths Jones, Harpreet Kaur Saini, Stijn van Dongen, and Anton J Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res*, 36(Database issue):D154–D158, 2008.

[41] David P Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.

[42] Ahmet M Denli, Bastiaan B J Tops, Ronald H A Plasterk, René F Ketting, and Gregory J Hannon. Processing of primary microRNAs by the Microprocessor complex. *Nature*, 432(7014):231–235, 2004.

[43] Jinju Han, Yoontae Lee, Kyu-Hyun Yeom, Young-Kook Kim, Hua Jin, and V. Narry Kim. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev*, 18(24):3016–3027, 2004.

[44] Richard I Gregory, Kai-Ping Yan, Govindasamy Amuthan, Thimmaiah Chendrimada, Behzad Doratotaj, Neil Cooch, and Ramin Shiekhattar. The Microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235–240, 2004.

[45] Markus Landthaler, Abdullah Yalcin, and Thomas Tuschl. The human DiGeorge syndrome critical region gene 8 and Its D. melanogaster homolog are required for miRNA biogenesis. *Curr Biol*, 14(23):2162–2167, 2004.

[46] Jinju Han, Yoontae Lee, Kyu-Hyeon Yeom, Jin-Wu Nam, Inha Heo, Je-Keun Rhee, Sun Young Sohn, Yunje Cho, Byoung-Tak Zhang, and V. Narry Kim. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, 125(5):887–901, 2006.

[47] J. Graham Ruby, Calvin H Jan, and David P Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–86, 2007.

[48] Katsutomo Okamura, Joshua W Hagen, Hong Duan, David M Tyler, and Eric C Lai. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell*, 130(1):89–100, 2007.

[49] Eugene Berezikov, Wei-Jen Chung, Jason Willis, Edwin Cuppen, and Eric C Lai. Mammalian mirtron genes. *Mol Cell*, 28(2):328–336, 2007.

[50] Rui Yi, Yi Qin, Ian G Macara, and Bryan R Cullen. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 17(24):3011–3016, 2003.

[51] Elsebet Lund, Stephan Güttinger, Angelo Calado, James E Dahlberg, and Ulrike Kutay. Nuclear export of microRNA precursors. *Science*, 303(5654):95–98, 2004.

[52] Yoontae Lee, Chiyoung Ahn, Jinju Han, Hyounjeong Choi, Jaekwang Kim, Jeongbin Yim, Junho Lee, Patrick Provost, Olof Rådmark, Sunyoung Kim, et al. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415–419, 2003.

[53] Klaus Förstemann, Yukihide Tomari, Tingting Du, Vasily V. Vagin, Ahmet M. Denli, Diana P. Bratu, Carla Klattenhoff, William E. Theurkauf, and Phillip D. Zamore. Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol*, 3(7):e236, 2005.

[54] Feng Jiang, Xuecheng Ye, Xiang Liu, Lauren Fincher, Dennis McKearin, and Qinghua Liu. Dicer-1 and R3D1-L catalyze microRNA maturation in Drosophila. *Genes Dev*, 19(14):1674–1679, 2005.

[55] Kuniaki Saito, Akira Ishizuka, Haruhiko Siomi, and Mikiko C Siomi. Processing of pre-microRNAs by the Dicer-1-Loquacious complex in Drosophila cells. *PLoS Biol*, 3(7):e235, 2005.

[56] Astrid D Haase, Lukasz Jaskiewicz, Haidi Zhang, Sébastien Lainé, Ragna Sack, Anne Gatignol, and Witold Filipowicz. TRBP, a regulator of cellular PKR and HIV-1 virus expression, interacts with Dicer and functions in RNA silencing. *EMBO Rep*, 6(10):961–967, 2005.

[57] Yoontae Lee, Inha Hur, Seong-Yeon Park, Young-Kook Kim, Mi Ra Suh, and V. Narry Kim. The role of PACT in the RNA silencing pathway. *EMBO J*, 25(3):522–532, 2006.

[58] Yukihide Tomari and Phillip D Zamore. Perspective: machines for RNAi. *Genes Dev*, 19(5):517–529, 2005.

[59] Anastasia Khvorova, Angela Reynolds, and Sumedha D Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–216, 2003.

[60] Yangming Wang, Rostislav Medvid, Collin Melton, Rudolf Jaenisch, and Robert Blelloch. DGCR8 is essential for microRNA biogenesis and silencing of embryonic stem cell self-renewal. *Nat Genet*, 39(3):380–385, 2007.

[61] Andrew Jakymiw, Shangli Lian, Theophany Eystathioy, Songqing Li, Minoru Satoh, John C Hamel, Marvin J Fritzler, and Edward K L Chan. Disruption of GW bodies impairs mammalian RNA interference. *Nat Cell Biol*, 7(12):1267–1274, 2005.

[62] Ana Eulalio, Eric Huntzinger, and Elisa Izaurralde. Getting to the root of miRNA-mediated gene silencing. *Cell*, 132(1):9–14, 2008.

[63] Xavier C Ding and Helge Grosshans. Repression of C. elegans microRNA targets at the initiation level of translation requires GW182 proteins. *EMBO J*, 28(3):213–222, 2009.

[64] Kyle Kai-How Farh, Andrew Grimson, Calvin Jan, Benjamin P Lewis, Wendy K Johnston, Lee P Lim, Christopher B Burge, and David P Bartel. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science*, 310(5755):1817–1821, 2005.

[65] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105, 2009.

[66] Soraya Yekta, I Hung Shih, and David P Bartel. MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, 304(5670):594–596, 2004.

[67] David P Bartel. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2):215–233, 2009.

[68] Dimos Gaidatzis, Erik van Nimwegen, Jean Hausser, and Mihaela Zavolan. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69, 2007.

[69] Andrew Grimson, Kyle Kai-How Farh, Wendy K Johnston, Philip Garrett Engele, Lee P Lim, and David P Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, 2007.

[70] John G Doench, Christian P Petersen, and Phillip A Sharp. siRNAs can function as miRNAs. *Genes Dev*, 17(4):438–442, 2003.

[71] Antonio J Giraldez, Ryan M Cinalli, Margaret E Glasner, Anton J Enright, J. Michael Thomson, Scott Baskerville, Scott M Hammond, David P Bartel, and Alexander F Schier. MicroRNAs regulate brain morphogenesis in zebrafish. *Science*, 308(5723):833–838, 2005.

[72] Antonio J Giraldez, Yuichiro Mishima, Jason Rihel, Russell J Grocock, Stijn Van Dongen, Kunio Inoue, Anton J Enright, and Alexander F Schier. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, 312(5770):75–79, 2006.

[73] Julius Brennecke, David R Hipfner, Alexander Stark, Robert B Russell, and Stephen M Cohen. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, 113(1):25–36, 2003.

[74] Peter S Linsley, Janell Schelter, Julja Burchard, Miho Kibukawa, Melissa M Martin, Steven R Bartz, Jason M Johnson, Jordan M Cummins, Christopher K Raymond, Hongyue Dai, et al. Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol Cell Biol*, 27(6):2240–2252, 2007.

[75] Peizhang Xu, Stephanie Y Vernooy, Ming Guo, and Bruce A Hay. The Drosophila microRNA Mir-14 suppresses cell death and is required for normal fat metabolism. *Curr Biol*, 13(9):790–795, 2003.

[76] Matthew N Poy, Lena Eliasson, Jan Krutzfeldt, Satoru Kuwajima, Xiaosong Ma, Patrick E Macdonald, Sébastien Pfeffer, Thomas Tuschl, Nikolaus Rajewsky, Patrik Rorsman, et al. A pancreatic islet-specific microRNA regulates insulin secretion. *Nature*, 432(7014):226–230, 2004.

[77] Jan Krützfeldt, Nikolaus Rajewsky, Ravi Braich, Kallanthottathil G. Rajeev, Thomas Tuschl, Muthiah Manoharan, and Markus Stoffel. Silencing of microRNAs in vivo with 'antagomirs'. *Nature*, 438(7068):685–689, 2005.

[78] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez Saavedra, Justin Lamb, David Peck, Alejandro Sweet Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, et al. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838, 2005.

[79] Aurora Esquela Kerscher and Frank J Slack. Oncomirs – microRNAs with a role in cancer. *Nat Rev Cancer*, 6(4):259–269, 2006.

[80] Marina Chekulaeva and Witold Filipowicz. Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. *Curr Opin Cell Biol*, 21(3):452–460, 2009.

[81] Lee P Lim, Nelson C Lau, Philip Garrett Engele, Andrew Grimson, Janell M Schelter, John Castle, David P Bartel, Peter S Linsley, and Jason M Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, 2005.

[82] Isabelle Behm Ansmant, Jan Rehwinkel, Tobias Doerks, Alexander Stark, Peer Bork, and Elisa Izaurralde. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev*, 20(14):1885–1898, 2006.

[83] Motoaki Wakiyama, Koji Takimoto, Osamu Ohara, and Shigeyuki Yokoyama. Let-7 microRNA-mediated mRNA deadenylation and translational repression in a mammalian cell-free system. *Genes Dev*, 21(15):1857–1862, 2007.

[84] Matthias Selbach, Björn Schwanhäusser, Nadine Thierfelder, Zhuo Fang, Raya Khanin, and Nikolaus Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, 2008.

[85] Daehyun Baek, Judit Villén, Chanseok Shin, Fernando D Camargo, Steven P Gygi, and David P Bartel. The impact of microRNAs on protein output. *Nature*, 455(7209):64–71, 2008.

[86] Young Sik Lee, Kenji Nakahara, John W Pham, Kevin Kim, Zhengying He, Erik J Sontheimer, and Richard W Carthew. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, 117(1):69–81, 2004.

[87] Katsutomo Okamura, Akira Ishizuka, Haruhiko Siomi, and Mikiko C Siomi. Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes Dev*, 18(14):1655–1666, 2004.

[88] H. Tabara, M. Sarkissian, W. G. Kelly, J. Fleenor, A. Grishok, L. Timmons, A. Fire, and C. C. Mello. The rde-1 gene, RNA interference, and transposon silencing in C. elegans. *Cell*, 99(2):123–132, 1999.

[89] Hiroaki Tabara, Erbay Yigit, Haruhiko Siomi, and Craig C Mello. The dsRNA binding protein RDE-4 interacts with RDE-1, DCR-1, and a DExH-box helicase to direct RNAi in C. elegans. *Cell*, 109(7):861–871, 2002.

[90] Shou-Wei Ding and Olivier Voinnet. Antiviral immunity directed by small RNAs. *Cell*, 130(3):413–426, 2007.

[91] Franck Vazquez. Arabidopsis endogenous small RNAs: highways and byways. *Trends Plant Sci*, 11(9):460–468, 2006.

[92] Caterina Catalanotto, Gianluca Azzalin, Giuseppe Macino, and Carlo Cogoni. Involvement of small RNAs and role of the qde genes in the gene silencing pathway in Neurospora. *Genes Dev*, 16(7):790–795, 2002.

[93] Brenda J Reinhart and David P Bartel. Small RNAs correspond to centromere heterochromatic repeats. *Science*, 297(5588):1831, 2002.

[94] Titia Sijen, Florian A Steiner, Karen L Thijssen, and Ronald H A Plasterk. Secondary siRNAs result from unprimed RNA synthesis and form a distinct class. *Science*, 315(5809):244–247, 2007.

[95] Julia Pak and Andrew Fire. Distinct populations of primary and secondary effectors during RNAi in C. elegans. *Science*, 315(5809):241–244, 2007.

[96] J. Graham Ruby, Calvin Jan, Christopher Player, Michael J Axtell, William Lee, Chad Nusbaum, Hui Ge, and David P Bartel. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, 127(6):1193–1207, 2006.

[97] Alexei A Aravin, Gregory J Hannon, and Julius Brennecke. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science*, 318(5851):761–764, 2007.

[98] Megha Ghildiyal, Hervé Seitz, Michael D Horwich, Chengjian Li, Tingting Du, Soohyun Lee, Jia Xu, Ellen L W Kittler, Maria L Zapp, Zhiping Weng, et al. Endogenous siRNAs derived from transposons and mRNAs in Drosophila somatic cells. *Science*, 320(5879):1077–1081, 2008.

[99] Katsutomo Okamura, Wei-Jen Chung, J. Graham Ruby, Huili Guo, David P Bartel, and Eric C Lai. The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, 453(7196):803–806, 2008.

[100] Katsutomo Okamura, Sudha Balla, Raquel Martin, Na Liu, and Eric C Lai. Two distinct mechanisms generate endogenous siRNAs from bidirectional transcription in Drosophila melanogaster. *Nat Struct Mol Biol*, 15(6):581–590, 2008.

[101] Yoshinori Kawamura, Kuniaki Saito, Taishin Kin, Yukiteru Ono, Kiyoshi Asai, Takafumi Sunohara, Tomoko N Okada, Mikiko C Siomi, and Haruhiko Siomi. Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, 453(7196):793–797, 2008.

[102] Benjamin Czech, Colin D Malone, Rui Zhou, Alexander Stark, Catherine Schlingeheyde, Monica Dus, Norbert Perrimon, Manolis Kellis, James A Wohlschlegel, Ravi Sachidanandam, et al. An endogenous small interfering RNA pathway in Drosophila. *Nature*, 453(7196):798–802, 2008.

[103] Wei-Jen Chung, Katsutomo Okamura, Raquel Martin, and Eric C Lai. Endogenous RNA interference provides a somatic defense against Drosophila transposons. *Curr Biol*, 18(11):795–802, 2008.

[104] Rui Zhou, Benjamin Czech, Julius Brennecke, Ravi Sachidanandam, James A Wohlschlegel, Norbert Perrimon, and Gregory J Hannon. Processing of Drosophila endo-siRNAs depends on a specific Loquacious isoform. *RNA*, 15(10):1886–1895, 2009.

[105] Julia Verena Hartig, Stephanie Esslinger, Romy Böttcher, Kuniaki Saito, and Klaus Förstemann. Endo-siRNAs depend on a new isoform of loquacious and target artificially introduced, high-copy sequences. *EMBO J*, 28(19):2932–2944, 2009.

[106] Toshiaki Watanabe, Atsushi Takeda, Tomoyuki Tsukiyama, Kazuyuki Mise, Tetsuro Okuno, Hiroyuki Sasaki, Naojiro Minami, and Hiroshi Imai. Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes. *Genes Dev*, 20(13):1732–1743, 2006.

[107] Oliver H Tam, Alexei A Aravin, Paula Stein, Angelique Girard, Elizabeth P Murchison, Sihem Cheloufi, Emily Hodges, Martin Anger, Ravi Sachidanandam, Richard M Schultz, et al. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, 453(7194):534–538, 2008.

[108] Toshiaki Watanabe, Yasushi Totoki, Atsushi Toyoda, Masahiro Kaneda, Satomi Kuramochi Miyagawa, Yayoi Obata, Hatsune Chiba, Yuji Kohara, Tomohiro Kono, Toru Nakano, et al. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, 453(7194):539–543, 2008.

[109] Joshua E Babiarz, J. Graham Ruby, Yangming Wang, David P Bartel, and Robert Blelloch. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev*, 22(20):2773–2785, 2008.

[110] J. Mauro Calabrese and Phillip A Sharp. Characterization of the short RNAs bound by the P19 suppressor of RNA silencing in mouse embryonic stem cells. *RNA*, 12(12):2092–2102, 2006.

[111] J. Mauro Calabrese, Amy C Seila, Gene W Yeo, and Phillip A Sharp. RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc Natl Acad Sci U S A*, 104(46):18097–18102, 2007.

[112] A. A. Aravin, N. M. Naumova, A. V. Tulin, V. V. Vagin, Y. M. Rozovsky, and V. A. Gvozdev. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr Biol*, 11(13):1017–1027, 2001.

[113] Alexei Aravin, Dimos Gaidatzis, Sébastien Pfeffer, Mariana Lagos Quintana, Pablo Landgraf, Nicola Iovino, Patricia Morris, Michael J Brownstein, Satomi Kuramochi Miyagawa, Toru Nakano, et al. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature*, 442(7099):203–207, 2006.

[114] Angélique Girard, Ravi Sachidanandam, Gregory J Hannon, and Michelle A Carmell. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 442(7099):199–202, 2006.

[115] Shane T Grivna, Ergin Beyret, Zhong Wang, and Haifan Lin. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev*, 20(13):1709–1714, 2006.

[116] Nelson C Lau, Anita G Seto, Jinkuk Kim, Satomi Kuramochi Miyagawa, Toru Nakano, David P Bartel, and Robert E Kingston. Characterization of the piRNA complex from rat testes. *Science*, 313(5785):363–367, 2006.

[117] Elizabeth P Murchison, Pouya Kheradpour, Ravi Sachidanandam, Carly Smith, Emily Hodges, Zhenyu Xuan, Manolis Kellis, Frank Grützner, Alexander Stark, and Gregory J Hannon. Conservation of small RNA pathways in platypus. *Genome Res*, 18(6):995–1004, 2008.

[118] Eric J Devor, Lingyan Huang, and Paul B Samollow. PiRNA-like RNAs in the marsupial Monodelphis domestica identify transcription clusters and likely marsupial transposon targets. *Mamm Genome*, 19(7-8):581–586, 2008.

[119] Saskia Houwing, Leonie M Kamminga, Eugene Berezikov, Daniela Cronembold, Angélique Girard, Hans van den Elst, Dmitri V Filippov, Heiko Blaser, Erez Raz, Cecilia B Moens, et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell*, 129(1):69–82, 2007.

[120] Saskia Houwing, Eugene Berezikov, and René F Ketting. Zili is required for germ cell differentiation and meiosis in zebrafish. *EMBO J*, 27(20):2702–2711, 2008.

[121] Anna Wilczynska, Nicola Minshall, Javier Armisen, Eric A Miska, and Nancy Standart. Two Piwi proteins, Xiwi and Xili, are expressed in the Xenopus female germline. *RNA*, 15(2):337–345, 2009.

[122] Nelson C Lau, Toshiro Ohsumi, Mark Borowsky, Robert E Kingston, and Michael D Blower. Systematic and single cell analysis of Xenopus Piwi-interacting RNAs and Xiwi. *EMBO J*, 28(19):2945–2958, 2009.

[123] Alexei A Aravin, Mariana Lagos Quintana, Abdullah Yalcin, Mihaela Zavolan, Debora Marks, Ben Snyder, Terry Gaasterland, Jutta Meyer, and Thomas Tuschl. The small RNA profile during Drosophila melanogaster development. *Dev Cell*, 5(2):337–350, 2003.

[124] Vasily V Vagin, Alla Sigova, Chengjian Li, Hervé Seitz, Vladimir Gvozdev, and Phillip D Zamore. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science*, 313(5785):320–324, 2006.

[125] Shinpei Kawaoka, Nobumitsu Hayashi, Susumu Katsuma, Hirohisa Kishino, Yuji Kohara, Kazuei Mita, and Toru Shimada. Bombyx small RNAs: genomic defense system against transposons in the silkworm, Bombyx mori. *Insect Biochem Mol Biol*, 38(12):1058–1065, 2008.

[126] Shinpei Kawaoka, Nobumitsu Hayashi, Yutaka Suzuki, Hiroaki Abe, Sumio Sugano, Yukihide Tomari, Toru Shimada, and Susumu Katsuma. The Bombyx ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. *RNA*, 15(7):1258–1264, 2009.

[127] Guilin Wang and Valerie Reinke. A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. *Curr Biol*, 18(12):861–867, 2008.

[128] Pedro J Batista, J. Graham Ruby, Julie M Claycomb, Rosaria Chiang, Noah Fahlgren, Kristin D Kasschau, Daniel A Chaves, Weifeng Gu, Jessica J Vasale, Shenghua Duan, et al. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Mol Cell*, 31(1):67–78, 2008.

[129] Dasaradhi Palakodeti, Magda Smielewska, Yi-Chien Lu, Gene W Yeo, and Brenton R Graveley. The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. *RNA*, 14(6):1174–1186, 2008.

[130] Marc R. Friedländer, Catherine Adamidi, Ting Han, Svetlana Lebedeva, Thomas A. Isenbarger, Martin Hirst, Marco Marra, Chad Nusbaum, William L. Lee, James C. Jenkin, et al. High-resolution profiling and discovery of planarian small RNAs. *Proc Natl Acad Sci U S A*, 106(28):11546–11551, 2009.

[131] Heather B Megosh, Daniel N Cox, Chris Campbell, and Haifan Lin. The role of PIWI and the miRNA machinery in Drosophila germline determination. *Curr Biol*, 16(19):1884–1894, 2006.

[132] 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(23):3715–3727, 1998.

[133] A. N. Harris and P. M. Macdonald. Aubergine encodes a Drosophila polar granule component required for pole cell formation and related to eIF2C. *Development*, 128(14):2823–2832, 2001.

[134] Julius Brennecke, Alexei A Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J Hannon. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell*, 128(6):1089–1103, 2007.

[135] Chengjian Li, Vasily V Vagin, Soohyun Lee, Jia Xu, Shengmei Ma, Hualin Xi, Hervé Seitz, Michael D Horwich, Monika Syrzycka, Barry M Honda, et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell*, 137(3):509–521, 2009.

[136] S. Kuramochi Miyagawa, T. Kimura, K. Yomogida, A. Kuroiwa, Y. Tadokoro, Y. Fujita, M. Sato, Y. Matsuda, and T. Nakano. Two mouse piwi-related genes: miwi and mili. *Mech Dev*, 108(1-2):121–133, 2001.

[137] Wei Deng and Haifan Lin. miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. *Dev Cell*, 2(6):819–830, 2002.

[138] Satomi Kuramochi Miyagawa, Tohru Kimura, Takashi W Ijiri, Taku Isobe, Noriko Asada, Yukiko Fujita, Masahito Ikawa, Naomi Iwai, Masaru Okabe, Wei Deng, et al. Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development*, 131(4):839–849, 2004.

[139] Michelle A Carmell, Angélique Girard, Henk J G van de Kant, Deborah Bourc'his, Timothy H Bestor, Dirk G de Rooij, and Gregory J Hannon. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell*, 12(4):503–514, 2007.

[140] Yingdee Unhavaithaya, Yi Hao, Ergin Beyret, Hang Yin, Satomi Kuramochi Miyagawa, Toru Nakano, and Haifan Lin. MILI, a PIWI-interacting RNA-binding protein, is required for germ line stem cell self-renewal and appears to positively regulate translation. *J Biol Chem*, 284(10):6507–6519, 2009.

[141] Alexei A Aravin, Ravi Sachidanandam, Deborah Bourc'his, Christopher Schaefer, Dubravka Pezic, Katalin Fejes Toth, Timothy Bestor, and Gregory J Hannon. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell*, 31(6):785–799, 2008.

[142] Satomi Kuramochi Miyagawa, Toshiaki Watanabe, Kengo Gotoh, Yasushi Totoki, Atsushi Toyoda, Masahito Ikawa, Noriko Asada, Kanako Kojima, Yuka Yamaguchi, Takashi W Ijiri, et al. DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes. *Genes Dev*, 22(7):908–917, 2008.

[143] Yohei Kirino and Zissimos Mourelatos. Mouse Piwi-interacting RNAs are 2'-O-methylated at their 3' termini. *Nat Struct Mol Biol*, 14(4):347–348, 2007.

[144] Tomoya Ohara, Yuriko Sakaguchi, Takeo Suzuki, Hiroki Ueda, Kenjyo Miyauchi, and Tsutomu Suzuki. The 3' termini of mouse Piwi-interacting RNAs are 2'-O-methylated. *Nat Struct Mol Biol*, 14(4):349–350, 2007.

[145] Kuniaki Saito, Yuriko Sakaguchi, Takeo Suzuki, Tsutomu Suzuki, Haruhiko Siomi, and Mikiko C Siomi. Pimet, the Drosophila homolog of HEN1, mediates 2'-O-methylation of Piwi- interacting RNAs at their 3' ends. *Genes Dev*, 21(13):1603–1608, 2007.

[146] Michael D Horwich, Chengjian Li, Christian Matranga, Vasily Vagin, Gwen Farley, Peng Wang, and Phillip D Zamore. The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol*, 17(14):1265–1272, 2007.

[147] Yohei Kirino and Zissimos Mourelatos. The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. *RNA*, 13(9):1397–1401, 2007.

[148] Ai Khim Lim and Toshie Kai. Unique germ-line organelle, nuage, functions to repress selfish genetic elements in Drosophila melanogaster. *Proc Natl Acad Sci U S A*, 104(16):6714–6719, 2007.

[149] Heather A Cook, Birgit S Koppetsch, Jing Wu, and William E Theurkauf. The Drosophila SDE3 homolog armitage is required for oskar mRNA silencing and embryonic axis specification. *Cell*, 116(6):817–829, 2004.

[150] Attilio Pane, Kristina Wehr, and Trudi Schüpbach. zucchini and squash encode two putative nucleases required for rasiRNA production in the Drosophila germline. *Dev Cell*, 12(6):851–862, 2007.

[151] Seth D Findley, Mio Tamanaha, Nigel J Clegg, and Hannele Ruohola Baker. Maelstrom, a Drosophila spindle-class gene, encodes a protein that colocalizes with Vasa and RDE1/AGO1 homolog, Aubergine, in nuage. *Development*, 130(5):859–871, 2003.

[152] Yu Chen, Attilio Pane, and Trudi Schüpbach. Cutoff and aubergine mutations result in retrotransposon upregulation and checkpoint activation in Drosophila. *Curr Biol*, 17(7):637–642, 2007.

[153] Alexei A Aravin, Mikhail S Klenov, Vasilii V Vagin, Frédéric Bantignies, Giacomo Cavalli, and Vladimir A Gvozdev. Dissection of a natural RNA silencing process in the Drosophila melanogaster germ line. *Mol Cell Biol*, 24(15):6742–6750, 2004.

[154] Alexei A Aravin, Ravi Sachidanandam, Angelique Girard, Katalin Fejes Toth, and Gregory J Hannon. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science*, 316(5825):744–747, 2007.

[155] N. Prud'homme, M. Gans, M. Masson, C. Terzian, and A. Bucheton. Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. *Genetics*, 139(2):697–711, 1995.

[156] Emeline Sarot, Geneviéve Payen Groschêne, Alain Bucheton, and Alain Pélisson. Evidence for a piwi-dependent RNA silencing of the gypsy endogenous retrovirus by the Drosophila melanogaster flamenco gene. *Genetics*, 166(3):1313–1321, 2004.

[157] Maryvonne Mével Ninio, Alain Pelisson, Jennifer Kinder, Ana Regina Campos, and Alain Bucheton. The flamenco locus controls the gypsy and ZAM retroviruses and is required for Drosophila oogenesis. *Genetics*, 175(4):1615–1624, 2007.

[158] Colin D Malone, Julius Brennecke, Monica Dus, Alexander Stark, W. Richard McCombie, Ravi Sachidanandam, and Gregory J Hannon. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell*, 137(3):522–535, 2009.

[159] Partha P Das, Marloes P Bagijn, Leonard D Goldstein, Julie R Woolford, Nicolas J Lehrbach, Alexandra Sapetschnig, Heeran R Buhecha, Michael J Gilchrist, Kevin L Howe, Rory Stark, et al. Piwi and piRNAs act upstream of an endogenous siRNA pathway to suppress Tc3 transposon mobility in the Caenorhabditis elegans germline. *Mol Cell*, 31(1):79–90, 2008.

[160] Raquel Assis and Alexey S Kondrashov. Rapid repetitive element-mediated expansion of piRNA clusters in mammalian evolution. *Proc Natl Acad Sci U S A*, 106(17):7079–7082, 2009.

[161] Julius Brennecke, Colin D Malone, Alexei A Aravin, Ravi Sachidanandam, Alexander Stark, and Gregory J Hannon. An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science*, 322(5906):1387–1392, 2008.

[162] Doron Betel, Robert Sheridan, Debora S Marks, and Chris Sander. Computational analysis of mouse piRNA sequence and biogenesis. *PLoS Comput Biol*, 3(11):e222, 2007.

[163] Shane T Grivna, Brook Pyhtila, and Haifan Lin. MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. *Proc Natl Acad Sci U S A*, 103(36):13415–13420, 2006.

[164] 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(1):25–33, 2000.

[165] Thomas A Volpe, Catherine Kidner, Ira M Hall, Grace Teng, Shiv I S Grewal, and Robert A Martienssen. Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi. *Science*, 297(5588):1833–1837, 2002.

[166] Titia Sijen and Ronald H A Plasterk. Transposon silencing in the Caenorhabditis elegans germ line by natural RNAi. *Nature*, 426(6964):310–314, 2003.

[167] 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(2):671–680, 1999.

[168] Sébastien Pfeffer, Mihaela Zavolan, Friedrich A Grässer, Minchen Chien, James J Russo, Jingyue Ju, Bino John, Anton J Enright, Debora Marks, Chris Sander, et al. Identification of virus-encoded microRNAs. *Science*, 304(5671):734–736, 2004.

[169] Sébastien Pfeffer, Alain Sewer, Mariana Lagos Quintana, Robert Sheridan, Chris Sander, Friedrich A Grässer, Linda F van Dyk, C. Kiong Ho, Stewart Shuman, Minchen Chien, et al. Identification of microRNAs of the herpesvirus family. *Nat Methods*, 2(4):269–276, 2005.

[170] Pablo Landgraf, Mirabela Rusu, Robert Sheridan, Alain Sewer, Nicola Iovino, Alexei Aravin, Sébastien Pfeffer, Amanda Rice, Alice O Kamphorst, Markus Landthaler, et al. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, 129(7):1401–1414, 2007.

[171] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[172] Simon Bennett. Solexa Ltd. *Pharmacogenomics*, 5(4):433–438, 2004.

[173] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, 1990.

[174] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning DNA sequences. *J Comput Biol*, 7(1-2):203–214, 2000.

[175] Eugene Berezikov, Geert van Tetering, Mark Verheul, Jose van de Belt, Linda van Laake, Joost Vos, Robert Verloop, Marc van de Wetering, Victor Guryev, Shuji Takada, et al. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res*, 16(10):1289–1298, 2006.

[176] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, et al. The transcriptional landscape of the mammalian genome. *Science*, 309(5740):1559–1563, 2005.

[177] Hristo B Houbaviy, Michael F Murray, and Phillip A Sharp. Embryonic stem cell-specific MicroRNAs. *Dev Cell*, 5(2):351–358, 2003.

[178] Victor Ambros. The functions of animal microRNAs. *Nature*, 431(7006):350–355, 2004.

[179] Yong Zhao, Eva Samal, and Deepak Srivastava. Serum response factor regulates a muscle-specific microRNA that targets Hand2 during cardiogenesis. *Nature*, 436(7048):214–220, 2005.

[180] Changchun Xiao, Dinis Pedro Calado, Gunther Galler, To-Ha Thai, Heide Christine Patterson, Jing Wang, Nikolaus Rajewsky, Timothy P Bender, and Klaus Rajewsky. MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb. *Cell*, 131(1):146–159, 2007.

[181] Lin He, J. Michael Thomson, Michael T Hemann, Eva Hernando Monge, David Mu, Summer Goodson, Scott Powers, Carlos Cordon Cardo, Scott W Lowe, Gregory J Hannon, et al. A microRNA polycistron as a potential human oncogene. *Nature*, 435(7043):828–833, 2005.

[182] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, 2005.

[183] Xiaohui Xie, Jun Lu, E. J. Kulbokas, Todd R Golub, Vamsi Mootha, Kerstin Lindblad Toh, Eric S Lander, and Manolis Kellis. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, 434(7031):338–345, 2005.

[184] Marc Rehmsmeier, Peter Steffen, Matthias Hochsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, 2004.

[185] Dang Long, Rosalind Lee, Peter Williams, Chi Yu Chan, Victor Ambros, and Ye Ding. Potent effect of target structure on microRNA function. *Nat Struct Mol Biol*, 14(4):287–294, 2007.

[186] Michael Kertesz, Nicola Iovino, Ulrich Unnerstall, Ulrike Gaul, and Eran Segal. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–1284, 2007.

[187] Jim C Huang, Tomas Babak, Timothy W Corson, Gordon Chua, Sofia Khan, Brenda L Gallie, Timothy R Hughes, Benjamin J Blencowe, Brendan J Frey, and Quaid D Morris. Using expression profiling data to identify human microRNA targets. *Nat Methods*, 4(12):1045–1049, 2007.

[188] Vincenzo Alessandro Gennarino, Marco Sardiello, Raffaella Avellino, Nicola Meola, Vincenza Maselli, Santosh Anand, Luisa Cutillo, Andrea Ballabio, and Sandro Banfi. MicroRNA target prediction by expression analysis of host genes. *Genome Res*, 19(3):481–490, 2009.

[189] Cydney B Nielsen, Noam Shomron, Rickard Sandberg, Eran Hornstein, Jacob Kitzman, and Christopher B Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–1910, 2007.

[190] Alexander Stark, Julius Brennecke, Natascha Bushati, Robert B Russell, and Stephen M Cohen. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell*, 123(6):1133–1146, 2005.

[191] Azra Krek, Dominic Grün, Matthew N Poy, Rachel Wolf, Lauren Rosenberg, Eric J Epstein, Philip MacMenamin, Isabelle da Piedade, Kristin C Gunsalus, Markus Stoffel, et al. Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500, 2005.

[192] Markus Landthaler, Dimos Gaidatzis, Andrea Rothballer, Po Yu Chen, Steven Joseph Soll, Lana Dinic, Tolulope Ojo, Markus Hafner, Mihaela Zavolan, and Thomas Tuschl. Molecular characterization of human Argonaute-containing ribonucleoprotein complexes and their bound target mRNAs. *RNA*, 14(12):2580–2596, 2008.

[193] Ida Schomburg, Antje Chang, Christian Ebeling, Marion Gremse, Christian Heldt, Gregor Huhn, and Dietmar Schomburg. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res*, 32(Database issue):D431–D433, 2004.

[194] Jonathan Bard, Seung Y Rhee, and Michael Ashburner. An ontology for cell types. *Genome Biol*, 6(2):R21, 2005.

[195] Philipp Berninger, Dimos Gaidatzis, Erik van Nimwegen, and Mihaela Zavolan. Computational analysis of small RNA cloning data. *Methods*, 44(1):13–21, 2008.

[196] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 33(Database issue):D501–D504, 2005.

[197] Andrew I Su, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, Richard Soden, Mimi Hayakawa, Gabriel Kreiman, et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A*, 101(16):6062–6067, 2004.

[198] Eric C Lai. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–364, 2002.

[199] John G Doench and Phillip A Sharp. Specificity of microRNA target selection in translational repression. *Genes Dev*, 18(5):504–511, 2004.

[200] Nikolaus Rajewsky and Nicholas D Socci. Computational identification of microRNA targets. *Dev Biol*, 267(2):529–535, 2004.

[201] Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, 2005.

[202] William H Majoros and Uwe Ohler. Spatial preferences of microRNA targets in 3' untranslated regions. *BMC Genomics*, 8:152, 2007.

[203] Hakim Tafer, Stefan L Ameres, Gregor Obernosterer, Christoph A Gebeshuber, Renée Schroeder, Javier Martinez, and Ivo L Hofacker. The impact of target site accessibility on the design of effective siRNAs. *Nat Biotechnol*, 26(5):578–583, 2008.

[204] Cristian I Castillo Davis and Daniel L Hartl. GeneMerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–892, 2003.

[205] R. M. Eglen. Muscarinic receptor subtypes in neuronal and non-neuronal cholinergic function. *Auton Autacoid Pharmacol*, 26(3):219–233, 2006.

[206] Sam Griffiths Jones, Russell J Grocock, Stijn van Dongen, Alex Bateman, and Anton J Enright. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–D144, 2006.

[207] Po Yu Chen, Heiko Manninga, Krasimir Slanchev, Minchen Chien, James J Russo, Jingyue Ju, Robert Sheridan, Bino John, Debora S Marks, Dimos Gaidatzis, et al. The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Genes Dev*, 19(11):1288–1293, 2005.

[208] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, et al. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res*, 36(Database issue):D773–D779, 2008.

[209] International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, 2004.

[210] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915):520–562, 2002.

[211] Paul Havlak, Rui Chen, K. James Durbin, Amy Egan, Yanru Ren, Xing-Zhi Song, George M Weinstock, and Richard A Gibbs. The Atlas genome assembly system. *Genome Res*, 14(4):721–732, 2004.

[212] Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, 428:493–521, 2004.

[213] Kerstin Lindblad Toh, Claire M Wade, Tarjei S Mikkelsen, Elinor K Karlsson, David B Jaffe, Michael Kamal, Michele Clamp, Jean L Chang, Edward J Kulbokas, Michael C Zody, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438(7069):803–819, 2005.

[214] Bovine Genome Sequencing, Analysis Consortium, Christine G Elsik, Ross L Tellam, Kim C Worley, Richard A Gibbs, Donna M Muzny, George M Weinstock, David L Adelson, Evan E Eichler, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*, 324(5926):522–528, 2009.

[215] M Kasahara, K Naruse, S Sasaki, Y Nakatani, W Qu, B Ahsan, T Yamada, Y Nagayasu, K Doi, Y Kasai, et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 446:714–719, 2007.

[216] Tamberlyn Bieri, Darin Blasiar, Philip Ozersky, Igor Antoshechkin, Carol Bastiani, Payan Canaran, Juancarlos Chan, Nansheng Chen, Wen J Chen, Paul Davis, et al. WormBase: new content and better access. *Nucleic Acids Res*, 35(Database issue):D506–D510, 2007.

[217] Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, 450(7167):203–218, 2007.

[218] Kelsey C. Martin and Anne Ephrussi. mRNA localization: gene expression in the spatial dimension. *Cell*, 136(4):719–730, 2009.

[219] Melissa J. Moore and Nick J. Proudfoot. Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell*, 136(4):688–700, 2009.

[220] Nahum Sonenberg and Alan G. Hinnebusch. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. *Cell*, 136(4):731–745, 2009.

[221] Adrienne E. McKee, Emmanuel Minet, Charlene Stern, Shervin Riahi, Charles D. Stiles, and Pamela A. Silver. A genome-wide in situ hybridization map of RNA-binding proteins reveals anatomically restricted expression in the developing mouse brain. *BMC Dev Biol*, 5:14, 2005.

[222] Jack D. Keene. RNA regulons: coordination of post-transcriptional events. *Nat Rev Genet*, 8(7):533–543, 2007.

[223] S. A. Tenenbaum, C. C. Carson, P. J. Lager, and J. D. Keene. Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays. *Proc Natl Acad Sci U S A*, 97(26):14085–14090, 2000.

[224] Isabel López de Silanes, Ming Zhan, Ashish Lal, Xiaoling Yang, and Myriam Gorospe. Identification of a target RNA motif for RNA-binding protein HuR. *Proc Natl Acad Sci U S A*, 101(9):2987–2992, 2004.

[225] André P. Gerber, Stefan Luschnig, Mark A. Krasnow, Patrick O. Brown, and Daniel Herschlag. Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. *Proc Natl Acad Sci U S A*, 103(12):4487–4492, 2006.

[226] J. R. Greenberg. Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Res*, 6(2):715–732, 1979.

[227] A. J. Wagenmakers, R. J. Reinders, and W. J. van Venrooij. Cross-linking of mRNA to proteins by irradiation of intact cells with ultraviolet light. *Eur J Biochem*, 112(2):323–330, 1980.

[228] G. Dreyfuss, Y. D. Choi, and S. A. Adam. Characterization of heterogeneous nuclear RNA-protein complexes in vivo with monoclonal antibodies. *Mol Cell Biol*, 4(6):1104–1114, 1984.

[229] S. Mayrand, B. Setyono, J. R. Greenberg, and T. Pederson. Structure of nuclear ribonucleoprotein: identification of proteins in contact with poly(A)+ heterogeneous nuclear RNA in living HeLa cells. *J Cell Biol*, 90(2):380–384, 1981.

[230] Jernej Ule, Kirk B. Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B. Darnell. CLIP identifies Nova-regulated RNA networks in the brain. *Science*, 302(5648):1212–1215, 2003.

[231] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, 2008.

[232] Gene W. Yeo, Nicole G. Coufal, Tiffany Y. Liang, Grace E. Peng, Xiang-Dong Fu, and Fred H. Gage. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol*, 16(2):130–137, 2009.

[233] Jeremy R. Sanford, Xin Wang, Matthew Mort, Natalia Vanduyn, David N. Cooper, Sean D. Mooney, Howard J. Edenberg, and Yunlong Liu. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res*, 19(3):381–394, 2009.

[234] Sander Granneman, Grzegorz Kudla, Elisabeth Petfalski, and David Tollervey. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci U S A*, 106(24):9613–9618, 2009.

[235] Sonia Guil and Javier F. Càceres. The multifunctional RNA-binding protein hnRNP A1 is required for processing of miR-18a. *Nat Struct Mol Biol*, 14(7):591–596, 2007.

[236] Sung Wook Chi, Julie B Zang, Aldo Mele, and Robert B Darnell. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–486, 2009.

[237] Dimitrios G. Zisoulis, Michael T. Lovci, Melissa L. Wilbert, Kasey R. Hutt, Tiffany Y. Liang, Amy E. Pasquinelli, and Gene W. Yeo. Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans. *Nat Struct Mol Biol*, 17(2):173–179, 2010.

[238] Yohei Kirino and Zissimos Mourelatos. Site-specific crosslinking of human microRNPs to RNA targets. *RNA*, 14(10):2254–2259, 2008.

[239] K. M. Meisenheimer and T. H. Koch. Photocross-linking of nucleic acids to associated proteins. *Crit Rev Biochem Mol Biol*, 32(2):101–140, 1997.

[240] A. Favre, G. Moreno, M. O. Blondel, J. Kliber, F. Vinzens, and C. Salet. 4-Thiouridine photosensitized RNA-protein crosslinking in mammalian cells. *Biochem Biophys Res Commun*, 141(2):847–854, 1986.

[241] Xiaoqiang Wang, Juanita McLachlan, Phillip D. Zamore, and Traci M Tanaka Hall. Modular recognition of RNA by a human pumilio-homology domain. *Cell*, 110(4):501–512, 2002.

[242] Markus Hafner, Pablo Landgraf, Janos Ludwig, Amanda Rice, Tolulope Ojo, Carolina Lin, Daniel Holoch, Cindy Lim, and Thomas Tuschl. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods*, 44(1):3–12, 2008.

[243] Marvin Wickens, David S. Bernstein, Judith Kimble, and Roy Parker. A PUF family portrait: 3'UTR regulation as a way of life. *Trends Genet*, 18(3):150–157, 2002.

[244] Rahul Siddharthan, Eric D. Siggia, and Erik van Nimwegen. PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*, 1(7):e67, 2005.

[245] Alessia Galgano, Michael Forrer, Lukasz Jaskiewicz, Alexander Kanitz, Mihaela Zavolan, and André P. Gerber. Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system. *PLoS One*, 3(9):e3164, 2008.

[246] André Galarneau and Stéphane Richard. Target RNA motif and target mRNAs of the Quaking STAR protein. *Nat Struct Mol Biol*, 12(8):691–698, 2005.

[247] Carol Anne Chénard and Stéphane Richard. New implications for the QUAKING RNA binding protein in human disease. *J Neurosci Res*, 86(2):233–242, 2008.

[248] Joel K. Yisraeli. VICKZ proteins: a multi-talented family of regulatory RNA-binding proteins. *Biol Cell*, 97(1):87–96, Jan 2005.

[249] Benjamin Boyerinas, Sun-Mi Park, Noam Shomron, Mads M. Hedegaard, Jeppe Vinther, Jens S. Andersen, Christine Feig, Jinbo Xu, Christopher B. Burge, and Marcus E. Peter. Identification of let-7-regulated oncofetal genes. *Cancer Res*, 68(8):2587–2591, 2008.

[250] Euthymios Dimitriadis, Theoni Trangas, Stavros Milatos, Periklis G. Foukas, Ioannis Gioulbasanis, Nelly Courtis, Finn C. Nielsen, Nikos Pandis, Urania Dafni, Georgia Bardi, et al. Expression of oncofetal RNA-binding protein CRD-BP/IMP1 predicts clinical outcome in colon cancer. *Int J Cancer*, 121(3):486–494, 2007.

[251] Diabetes Genetics Initiative. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, 316(5829):1331–1336, 2007.

[252] Nikolaus Rajewsky. microRNA target predictions in animals. *Nat Genet*, 38 Suppl:S8–13, 2006.

[253] Yanli Wang, Stefan Juranek, Haitao Li, Gang Sheng, Greg S Wardle, Thomas Tuschl, and Dinshaw J Patel. Nucleation, propagation and cleavage of target RNAs in Ago silencing complexes. *Nature*, 461(7265):754–761, 2009.

[254] Jean Hausser, Markus Landthaler, Lukasz Jaskiewicz, Dimos Gaidatzis, and Mihaela Zavolan. Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C-miRNA complexes and the degradation of miRNA targets. *Genome Res*, 2009.

[255] George Easow, Aurelio A. Teleman, and Stephen M. Cohen. Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13(8):1198–1204, 2007.

[256] Joshua J Forman, Aster Legesse Miller, and Hilary A Coller. A search for conserved sequences in coding regions reveals that the let-7 microRNA targets Dicer within its coding sequence. *Proc Natl Acad Sci U S A*, 105(39):14879–14884, 2008.

[257] J. Robin Lytle, Therese A Yario, and Joan A Steitz. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. *Proc Natl Acad Sci U S A*, 104(23):9667–9672, 2007.

[258] Ulf Andersson Ørom, Finn Cilius Nielsen, and Anders H. Lund. MicroRNA-10a binds the 5'UTR of ribosomal protein mRNAs and enhances their translation. *Mol Cell*, 30(4):460–471, 2008.

[259] Yvonne Tay, Jinqiu Zhang, Andrew M Thomson, Bing Lim, and Isidore Rigoutsos. MicroRNAs to Nanog, Oct4 and Sox2 coding regions modulate embryonic stem cell differentiation. *Nature*, 455(7216):1124–1128, 2008.

[260] P. M. Sharp and W. H. Li. The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*, 15(3):1281–1295, 1987.

[261] Shuo Gu, Lan Jin, Feijie Zhang, Peter Sarnow, and Mark A Kay. Biological basis for restriction of microRNA targets to the 3' untranslated region in mammalian mRNAs. *Nat Struct Mol Biol*, 16(2):144–150, 2009.

[262] Sabbi Lall, Dominic Grün, Azra Krek, Kevin Chen, Yi-Lu Wang, Colin N. Dewey, Pranidhi Sood, Teresa Colombo, Nicolas Bray, Philip Macmenamin, et al. A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol*, 16(5):460–471, 2006.

[263] Bradley M. Lunde, Claire Moore, and Gabriele Varani. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*, 8(6):479–490, 2007.

[264] Monica C. Vella, Eun-Young Choi, Shin-Yi Lin, Kristy Reinert, and Frank J. Slack. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the lin-41 3'UTR. *Genes Dev*, 18(2):132–137, 2004.

[265] Gunter Meister and Thomas Tuschl. Mechanisms of gene silencing by double-stranded RNA. *Nature*, 431(7006):343–349, 2004.

[266] Lin He and Gregory J Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet*, 5(7):522–531, 2004.

[267] Phillip D Zamore and Benjamin Haley. Ribo-gnome: the big world of small RNAs. *Science*, 309(5740):1519–1524, 2005.

[268] V. Narry Kim. MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol*, 6(5):376–385, 2005.

[269] Natascha Bushati and Stephen M Cohen. microRNA functions. *Annu Rev Cell Dev Biol*, 23:175–205, 2007.

[270] Ramesh S Pillai, Suvendra N Bhattacharyya, and Witold Filipowicz. Repression of protein synthesis by miRNAs: how many mechanisms? *Trends Cell Biol*, 17(3):118–126, 2007.

[271] Marco Antonio Valencia Sanchez, Jidong Liu, Gregory J Hannon, and Roy Parker. Control of translation and mRNA degradation by miRNAs and siRNAs. *Genes Dev*, 20(5):515–524, 2006.

[272] A. G. Smith. Embryo-derived stem cells: of mice and men. *Annu Rev Cell Dev Biol*, 17:435–462, 2001.

[273] Laurie A Boyer, Tong Ihn Lee, Megan F Cole, Sarah E Johnstone, Stuart S Levine, Jacob P Zucker, Matthew G Guenther, Roshan M Kumar, Heather L Murray, Richard G Jenner, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, 122(6):947–956, 2005.

[274] Yuin-Han Loh, Qiang Wu, Joon-Lin Chew, Vinsensius B Vega, Weiwei Zhang, Xi Chen, Guillaume Bourque, Joshy George, Bernard Leong, Jun Liu, et al. The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet*, 38(4):431–440, 2006.

[275] Mi-Ra Suh, Yoontae Lee, Jung Yeon Kim, Soo-Kyoung Kim, Sung-Hwan Moon, Ji Yeon Lee, Kwang-Yul Cha, Hyung Min Chung, Hyun Soo Yoon, Shin Yong Moon, et al. Human embryonic stem cells express a unique set of microRNAs. *Dev Biol*, 270(2):488–498, 2004.

[276] Fuchou Tang, Masahiro Kaneda, Dónal O'Carroll, Petra Hajkova, Sheila C Barton, Y. Andrew Sun, Caroline Lee, Alexander Tarakhovsky, Kaiqin Lao, and M. Azim Surani. Maternal microRNAs are essential for mouse zygotic development. *Genes Dev*, 21(6):644–648, 2007.

[277] Chryssa Kanellopoulou, Stefan A Muljo, Andrew L Kung, Shridar Ganesan, Ronny Drapkin, Thomas Jenuwein, David M Livingston, and Klaus Rajewsky. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. *Genes Dev*, 19(4):489–501, 2005.

[278] Elizabeth P Murchison, Janet F Partridge, Oliver H Tam, Sihem Cheloufi, and Gregory J Hannon. Characterization of Dicer-deficient murine embryonic stem cells. *Proc Natl Acad Sci U S A*, 102(34):12135–12140, 2005.

[279] Daniela Schmitter, Jody Filkowski, Alain Sewer, Ramesh S Pillai, Edward J Oakeley, Mihaela Zavolan, Petr Svoboda, and Witold Filipowicz. Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res*, 34(17):4801–4815, 2006.

[280] Fanyi Zeng, Don A Baldwin, and Richard M Schultz. Transcript profiling during preimplantation mouse development. *Dev Biol*, 272(2):483–496, 2004.

[281] Taiping Chen, Yoshihide Ueda, Shaoping Xie, and En Li. A novel Dnmt3a isoform produced from an alternative promoter localizes to euchromatin and its expression correlates with active de novo methylation. *J Biol Chem*, 277(41):38746–38754, 2002.

[282] Ulla Aapola, Katja Mäenpää, Antti Kaipia, and Pärt Peterson. Epigenetic modifications affect Dnmt3L expression. *Biochem J*, 380(Pt 3):705–713, 2004.

[283] Artit Jinawath, Satoshi Miyake, Yuka Yanagisawa, Yoshimitsu Akiyama, and Yasuhito Yuasa. Transcriptional regulation of the human DNA methyltransferase 3A and 3B genes by Sp3 and Sp1 zinc finger proteins. *Biochem J*, 385(Pt 2):557–564, 2005.

[284] Andrew I Su, Michael P Cooke, Keith A Ching, Yaron Hakak, John R Walker, Tim Wiltshire, Anthony P Orth, Raquel G Vega, Lisa M Sapinoso, Aziz Moqrich, et al. Large-scale analysis of the human and mouse transcriptomes. *Proc Natl Acad Sci U S A*, 99(7):4465–4470, 2002.

[285] M. A. Harris, J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue):D258–D261, 2004.

[286] Larisa Litovchick, Subhashini Sadasivam, Laurence Florens, Xiaopeng Zhu, Selene K Swanson, Soundarapandian Velmurugan, Runsheng Chen, Michael P Washburn, X. Shirley Liu, and James A DeCaprio. Evolutionarily conserved multisubunit RBL2/p130 and E2F4 protein complex represses human cell cycle-dependent genes in quiescence. *Mol Cell*, 26(4):539–551, 2007.

[287] Peili Gu, Damien Le Menuet, Arthur C-K Chung, and Austin J Cooney. Differential recruitment of methylated CpG binding domains by the orphan receptor GCNF initiates the repression and silencing of Oct4 expression. *Mol Cell Biol*, 26(24):9471–9483, 2006.

[288] Nirit Feldman, Ariela Gerson, Jia Fang, En Li, Yi Zhang, Yoichi Shinkai, Howard Cedar, and Yehudit Bergman. G9a-mediated irreversible epigenetic inactivation of Oct-3/4 during early embryogenesis. *Nat Cell Biol*, 8(2):188–194, 2006.

[289] Jing-Yu Li, Min-Tie Pu, Ryutaro Hirasawa, Bin-Zhong Li, Yan-Nv Huang, Rong Zeng, Nai-He Jing, Taiping Chen, En Li, Hiroyuki Sasaki, et al. Synergistic function of DNA methyltransferases Dnmt3a and Dnmt3b in the methylation of Oct4 and Nanog. *Mol Cell Biol*, 27(24):8748–8759, 2007.

[290] Andrei L Gartel and Senthil K Radhakrishnan. Lost in transcription: p21 repression, mechanisms, and consequences. *Cancer Res*, 65(10):3980–3985, 2005.

[291] V. Nordhoff, K. Hübner, A. Bauer, I. Orlova, A. Malapetsa, and H. R. Schöler. Comparative analysis of human, bovine, and murine Oct-4 upstream promoter sequences. *Mamm Genome*, 12(4):309–317, 2001.

[292] Mirei Murakami, Tomoko Ichisaka, Mitsuyo Maeda, Noriko Oshiro, Kenta Hara, Frank Edenhofer, Hiroshi Kiyama, Kazuyoshi Yonezawa, and Shinya Yamanaka. mTOR is essential for growth and proliferation in early mouse embryos and embryonic stem cells. *Mol Cell Biol*, 24(15):6710–6718, 2004.

[293] Richard M Schultz. The molecular foundations of the maternal to zygotic transition in the preimplantation embryo. *Hum Reprod Update*, 8(4):323–331, 2002.

[294] Wen-Yee Choi, Antonio J Giraldez, and Alexander F Schier. Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. *Science*, 318(5848):271–274, 2007.

[295] G. P. Dotto. p21(WAF1/Cip1): more than a break to the cell cycle? *Biochim Biophys Acta*, 1471(1):M43–M56, 2000.

[296] Ilona Zvetkova, Anwyn Apedaile, Bernard Ramsahoye, Jacqueline E Mermoud, Lucy A Crompton, Rosalind John, Robert Feil, and Neil Brockdorff. Global hypomethylation of the genome in XX embryonic stem cells. *Nat Genet*, 37(11):1274–1279, 2005.

[297] Muller Fabbri, Ramiro Garzon, Amelia Cimmino, Zhongfa Liu, Nicola Zanesi, Elisa Callegari, Shujun Liu, Hansjuerg Alder, Stefan Costinean, Cecilia Fernandez Cymering, et al. MicroRNA-29 family reverts aberrant methylation in lung cancer by targeting DNA methyltransferases 3A and 3B. *Proc Natl Acad Sci U S A*, 104(40):15805–15810, 2007.

[298] Taiping Chen, Yoshihide Ueda, Jonathan E Dodge, Zhenjuan Wang, and En Li. Establishment and maintenance of genomic methylation patterns in mouse embryonic stem cells by Dnmt3a and Dnmt3b. *Mol Cell Biol*, 23(16):5594–5605, 2003.

[299] Chisaki Ishida, Kiyoe Ura, Akiko Hirao, Hiroyuki Sasaki, Atsushi Toyoda, Yoshiyuki Sakaki, Hitoshi Niwa, En Li, and Yasufumi Kaneda. Genomic organization and promoter analysis of the Dnmt3b gene. *Gene*, 310:151–159, 2003.

[300] Keisuke Nimura, Chisaki Ishida, Hiroshi Koriyama, Kenichiro Hata, Shinya Yamanaka, En Li, Kiyoe Ura, and Yasufumi Kaneda. Dnmt3a2 targets endogenous Dnmt3L to ES cell chromatin and induces regional DNA methylation. *Genes Cells*, 11(10):1225–1237, 2006.

[301] Lasse Sinkkonen, Marjo Malinen, Katri Saavalainen, Sami Väisänen, and Carsten Carlberg. Regulation of the human cyclin C gene via multiple vitamin D3-responsive regions in its promoter. *Nucleic Acids Res*, 33(8):2440–2451, 2005.

[302] Noora Kotaja, Suvendra N Bhattacharyya, Lukasz Jaskiewicz, Sarah Kimmins, Martti Parvinen, Witold Filipowicz, and Paolo Sassone Corsi. The chromatoid body of male germ cells: similarity with processing bodies and presence of Dicer and microRNA pathway components. *Proc Natl Acad Sci U S A*, 103(8):2647–2652, 2006.

[303] Zhijin Wu, Rafael A. Irizarry, Robert Gentleman, Frabcisco Martinez Murillo, and Forrest Spencer. A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association*, 99:909–917, 2004.

[304] K. D. Pruitt, K. S. Katz, H. Sicotte, and D. R. Maglott. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet*, 16(1):44–47, 2000.

[305] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3:Article3, 2004.

[306] Nayoung Suh, Lauren Baehner, Felix Moltzahn, Collin Melton, Archana Shenoy, Jing Chen, and Robert Blelloch. MicroRNA function is globally suppressed in mouse oocytes and early embryos. *Curr Biol*, 20(3):271–277, 2010.

[307] Lasse Sinkkonen, Tabea Hugenschmidt, Philipp Berninger, Dimos Gaidatzis, Fabio Mohn, Caroline G Artus Revel, Mihaela Zavolan, Petr Svoboda, and Witold Filipowicz. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. *Nat Struct Mol Biol*, 15(3):259–267, 2008.

[308] Elizabeth P Murchison, Paula Stein, Zhenyu Xuan, Hua Pan, Michael Q Zhang, Richard M Schultz, and Gregory J Hannon. Critical roles for Dicer in the female germline. *Genes Dev*, 21(6):682–693, 2007.

[309] You-Qiang Su, Koji Sugiura, Yong Woo, Karen Wigglesworth, Sonya Kamdar, Jason Affourtit, and John J Eppig. Selective degradation of transcripts during meiotic maturation of mouse oocytes. *Dev Biol*, 302(1):104–117, 2007.

[310] Stijn van Dongen, Cei Abreu Goodger, and Anton J Enright. Detecting microRNA binding and siRNA off-target effects from expression data. *Nat Methods*, 5(12):1023–1025, 2008.

[311] Masahiro Kaneda, Fuchou Tang, Dónal O'Carroll, Kaiqin Lao, and M. Azim Surani. Essential role for Argonaute2 protein in mouse oogenesis. *Epigenetics Chromatin*, 2(1):9, 2009.

[312] Mareike Puschendorf, Paula Stein, Edward J. Oakeley, Richard M. Schultz, Antoine H F M. Peters, and Petr Svoboda. Abundant transcripts from retrotransposons are unstable in fully grown mouse oocytes. *Biochem Biophys Res Commun*, 347(1):36–43, 2006.

[313] Jesse Salisbury, Keith W. Hutchison, Karen Wigglesworth, John J. Eppig, and Joel H. Graber. Probe-level analysis of expression microarrays characterizes isoform-specific degradation during mouse oocyte maturation. *PLoS One*, 4(10):e7479, 2009.

[314] Aimee L. Jackson, Julja Burchard, Janell Schelter, B Nelson Chau, Michele Cleary, Lee Lim, and Peter S. Linsley. Widespread siRNA "off-target" transcript silencing mediated by seed region sequence complementarity. *RNA*, 12(7):1179–1187, 2006.

[315] Sam E V. Linsen, Elzo de Wit, Georges Janssens, Sheila Heater, Laura Chapman, Rachael K. Parkin, Brian Fritz, Stacia K. Wyman, Ewart de Bruijn, Emile E. Voest, et al. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods*, 6(7):474–476, 2009.

[316] Jidong Liu, Marco Antonio Valencia Sanchez, Gregory J Hannon, and Roy Parker. MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nat Cell Biol*, 7(7):719–723, 2005.

[317] Ramesh S Pillai, Suvendra N Bhattacharyya, Caroline G Artus, Tabea Zoller, Nicolas Cougot, Eugenia Basyuk, Edouard Bertrand, and Witold Filipowicz. Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science*, 309(5740):1573–1576, 2005.

[318] Matyas Flemr, Jun Ma, Richard M. Schultz, and Petr Svoboda. P-body loss is concomitant with formation of a messenger RNA storage domain in mouse oocytes. *Biol Reprod*, 82(5):1008–1017, 2010.

[319] Preethi H Gunaratne. Embryonic Stem Cell MicroRNAs: Defining Factors in Induced Pluripotent (iPS) and Cancer (CSC) Stem Cells? *Curr Stem Cell Res Ther*, 2009.

[320] Alexander Marson, Stuart S. Levine, Megan F. Cole, Garrett M. Frampton, Tobias Brambrink, Sarah Johnstone, Matthew G. Guenther, Wendy K. Johnston, Marius Wernig, Jamie Newman, et al. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell*, 134(3):521–533, 2008.

[321] R. M. Schultz, R. R. Montgomery, and J. R. Belanoff. Regulation of mouse oocyte meiotic maturation: implication of a decrease in oocyte cAMP and protein dephosphorylation in commitment to resume meiosis. *Dev Biol*, 97(2):264–273, 1983.

[322] C. L. Chatot, C. A. Ziomek, B. D. Bavister, J. L. Lewis, and I. Torres. An improved culture medium supports development of random-bred 1-cell mouse embryos in vitro. *J Reprod Fertil*, 86(2):679–688, 1989.

[323] P. Svoboda, P. Stein, and R. M. Schultz. RNAi in mouse oocytes and preimplantation embryos: effectiveness of hairpin dsRNA. *Biochem Biophys Res Commun*, 287(5):1099–1104, 2001.

[324] S. Kurasawa, R. M. Schultz, and G. S. Kopf. Egg-induced modifications of the zona pellucida of mouse eggs: effects of microinjected inositol 1,4,5-trisphosphate. *Dev Biol*, 133(1):295–304, 1989.

[325] Hua Pan, Marilyn J. O'brien, Karen Wigglesworth, John J. Eppig, and Richard M. Schultz. Transcript profiling during mouse oocyte development and the effect of gonadotropin priming and development in vitro. *Dev Biol*, 286(2):493–506, 2005.

[326] Hang Yin and Haifan Lin. An epigenetic activation role of Piwi and a Piwi-associated piRNA in Drosophila melanogaster. *Nature*, 450(7167):304–308, 2007.

[327] Kuniaki Saito, Sachi Inagaki, Toutai Mituyama, Yoshinori Kawamura, Yukiteru Ono, Eri Sakota, Hazuki Kotani, Kiyoshi Asai, Haruhiko Siomi, and Mikiko C Siomi. A regulatory circuit for piwi by the large Maf gene traffic jam in Drosophila. *Nature*, 461(7268):1296–1299, 2009.

[328] A. J. Olson, J. Brennecke, A. A. Aravin, G. J. Hannon, and R. Sachidanandam. Analysis of large-scale sequencing of small RNAs. *Pac Symp Biocomput*, pages 126–136, 2008.

Curriculum Vitae

EDUCATION

May 06 - Dec 09	PhD Student in Bioinformatics at Biozentrum Basel in Mihaela Zavolans group, Switzerland
Sep 00 - Mar 06	Diploma in Computer Science with Minor in Bioinformatics, Julius-Maximilans University, Wuerzburg, Germany
Oct 99 - Aug 00	Civilian service at Seniorenresidenz, Woerth am Main, Germany
May 99	High school graduation, Julius-Echter Gymnasium, Elsenfeld, Germany

EXPERIENCE

Jan 10 - present	EMBL Outstation, Grenoble, France Postdoctoral researcher
Dec 05 - Mar 06	Chair for Bioinformatics, Wuerzburg, Germany Undergraduate research assistant
Apr 02 - Mar 03	Chair for Computer Science I, Wuerzburg, Germany Undergraduate teaching assistant

PUBLICATIONS

Berninger P, Jaskiewicz L, Zavolan M. Conserved generation of short products at piRNA loci, **BMC Genomics** 2011 Jan 19;12:46.

Hafner M*, Landthaler M*, Burger L, Khorshid M, Hausser J, **Berninger P**, Rothballer A, Ascano M, Jungkamp A, Munschauer M, Ulrich A, Wardle GS, Dewell S, Zavolan M, Tuschl T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP, **Cell**, 2010 Apr 2;141(1):129-41

Ma J, Flemr M, Stein P, **Berninger P**, Malik R, Zavolan M, Svoboda P, Schultz RM. MicroRNA activity is suppressed in mouse oocytes. **Curr Biol.** 2010 Feb 9;20(3):265-70.

Hausser J, **Berninger P**, Rodak C, Jantscher Y, Wirth S, Zavolan M. MirZ: an integrated microRNA expression atlas and target prediction resource. **Nucleic Acids Res.** 2009 May 25 (37).

Pena JT, Sohn-Lee C, Rouhanifard SH, Ludwig J, Hafner M, Mihailovic A, Lim C, Holoch D, **Berninger P**, Zavolan M, Tuschl T. miRNA in situ hybridization in formaldehyde and EDC-fixed tissues. **Nat Methods.** 2009 Feb;6(2):139-41

Sinkkonen L, Hugenschmidt T, **Berninger P**, Gaidatzis D, Mohn F, Artus-Revel CG, Zavolan M, Svoboda P, Filipowicz W. MicroRNAs control de novo DNA methylation through regulation of transcriptional repressors in mouse embryonic stem cells. **Nat Struct Mol Biol.** 2008 Mar;15(3):259-67.

Berninger P*, Gaidatzis D*, van Nimwegen E, Zavolan M. Computational analysis of small RNA cloning data. **Methods.** 2008 Jan;44(1):13-21

POSTERS

Berninger P, Jaskiewicz L, Khorshid M, Zavolan M. Computational study of piRNA sequence reads reveals short by-products of the ping-pong mechanism. 5^{th} Microsymposium on small RNAs, Vienna, Austria, 2010

Berninger P, Jaskiewicz L, Zavolan M. Reanalysis of piRNA sequence reads reveals short byproducts of the ping-pong cycle. Biocenter symposium, Basel, Switzerland, 2009

Berninger P*, Hausser J*, Khorshid M*, Hafner M, Landthaler M, Tuschl T, Zavolan M. Identification of miRNA targets. SIB Meeting, Fribourg,Switzerland, 2009.

Berninger P, Gaidatzis D, van Nimwegen E, Zavolan M. Computational analysis of small RNA cloning data. RNAi Europe, Barcelona, Spain, 2007, 5^{th} Colmar symposium: the new RNA frontiers, Colmar, France, 2007 & SIB Meeting, Grindelwald, Switzerland 2007

Berninger P, Gaidatzis D, Zavolan M. Differential Targeting of 3'UTRs. Regulatory RNAs in Eukaryotes, Mittelwihr, France, 2006.

CONFERENCES

5^{th} Microsymposium on small RNAs, Vienna, Austria, 2010

Basel Computational Biology Conference, Basel, Switzerland, 2009.

5^{th} Colmar symposium: the new RNA frontiers, Colmar, France, 2007.

RNAi Europe, Barcelona, Spain, 2007.

The Future of Genome Research in the Light of Ultrafast Sequencing Technologies, Bielefeld, Germany, 2007.

Basel Computational Biology Conference, Basel, Switzerland, 2007.

MicroRNAs and siRNAs: Biological Functions and Mechanisms, Keystone, USA, 2007.

Regulatory RNAs in Eukaryotes, Mittelwihr, France, 2006.

SKMB Gene Regulation workshop, Lausanne, Switzerland, 2006.

WORKSHOPS & TRAINING

Advanced RNA-Seq and ChiP-Seq Data Analysis, Hinxton, UK, 2011

Computational Analysis of UHT Sequencing Data, Lausanne, Switzerland, 2009.

Protein structure: prediction and analysis, Lausanne, Switzerland, 2008

EMBO Practical Course on Computational RNA Biology, Cargese, France, 2008.

FELLOWSHIPS

SNF Fellowship for prospective researchers, 2010