Discovery and profiling of animal small RNAs using deep sequencing

Inaugural-Dissertation
to obtain the academic degree
Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry and Pharmacy
of Freie Universität Berlin

by

Marc Riemer Friedländer

from København

2009

Time period of doctorate studies: June 1st 2006 – December 1st 2009.

Supervisor of doctorate studies: Prof. Nikolaus Rajewsky

Institute of doctorate studies: Max Delbrück Center for Molecular Medicine

Date of disputation defence: February 1st 2010

First reviewer:

Prof. Nikolaus Rajewsky
Division of Systems Biology
Max Delbrück Center for Molecular Medicine
D-13125 Berlin
Email: rajewsky@mdc-berlin.de
Phone + 30 9406 2999

Second reviewer:

Prof. Wolfgang Schuster
Applied Genetics
Freie Universität
D-14195 Berlin
Email: schuwowi@zedat.fu-berlin.de
Phone +30 838 56797

Acknowledgements

I would like to dedicate this work to my mother Hia and my girlfriend Petra who have waited patiently and faithfully for me while I have been away for three and a half years on this scientific Odyssey.

First I would like to give my warmest thanks to my supervisor Nikolaus Rajewsky, who has been a lighthouse of inspiration and has guided me on a straight course between the reefs of idleness and procrastination. Second, I would like to thank Azra Krek, with whom I started this journey and who has by now passed through the underworld of thesis writing and has reached her sunny shores of New York. Third, I would like give my fullest thanks to every single one of the stout hoplites and amazons on board the Rajewsky galley, especially to my navigator and planarian co-conspirator Catherine Adamidi and to my oarsman and fellow miRDeepie Sebastian Mackowiak without whose help I would certainly have sailed in circles. Fourth, I would like to thank our scientific collaborators and also Jennifer Stewart who helped me through the maelstrom of university bureaucracy. Fifth, I would like to thank the cat of the ship, Giaco, who bravely volunteered to donate blood when it looked like we could not obtain a dog lymphocyte sample. Last, I would like to thank the MDC fellowship which has kept me supplied with water and biscuits on the journey.

INTRODUCTION

Motivation

Discoveries in the last decade have shown that small RNAs (such as microRNAs) perform a number of important functions, including post-transcriptional gene regulation, transposon silencing, DNA methylation, chromatin modifications and chromosome segregation, e.g.[1-6]. The ability of the new deep sequencing technologies to sequence millions of short RNAs in a few hours have made them the method of choice for simultaneous discovery and profiling[i] of small RNAs[7, 8]. However, when the sequenced RNAs are mapped to the reference genome, they typically locate to millions of distinct loci, only a few of which are loci that produce regulatory small RNAs. To distinguish the few loci that produce regulatory small RNAs from the many loci that are sources of other short RNAs like degradation products is a non-trivial computational challenge. In my doctorate work, I have attempted formalize knowledge of small RNA biology into computational models that can be used to discover and profile deep sequenced small RNAs.

Animal and plant small RNAs

There is emerging evidence that regulatory small RNAs are present in bacteria as well as in eukaryotes[9, 10]. However, the full ensemble of small RNA interacting proteins that enacts canonical RNA interference is found only in the eukaryotic clades animals, plants and fungi. It is possible that RNA interference has developed in early eukaryotes as a defense mechanism that cleaves double-stranded RNA viruses into harmless short RNAs[11]. Later in evolution, this defense mechanism may have been adapted to also cleave endogenous double-stranded RNA for various regulatory uses[12]. Consistent with the hypothesis that regulatory small RNAs have largely emerged in animals and plants through convergent evolution, animal and plant small RNAs display different characteristics[13]. The scope of this thesis is limited to the animal small RNAs. More specifically, it will focus on the two most studied classes of animal small RNAs: miRNAs (microRNAs) and piRNAs.

[i] 'Profiling' includes expression profiling, but also other small RNA readouts that can be obtained from sequencing such as length distributions, variations in 5' ends and 3' ends etc.

miRNAs

miRNAs are regulatory small RNAs ~22 nts in length that are bound by the miRNP protein complex[14, 15]. miRNAs guide the complex to target sites in the 3'UTRs or, rarely, the coding sequence of mRNAs, causing mRNA degradation or translational to be inhibited[1, 16-20]. Thus, miRNAs reduce and/or buffer the expression of protein coding genes. All metazoan animals investigated have miRNA genes, ranging in number from ~40 (sea anemone) to ~700 (humans)[21, 22]. miRNA genes appear to be constantly gained throughout evolution, thus some are deeply conserved and some are species-specific [23-26]. Many miRNAs target hundreds of mRNAs, and it is estimated that between 30% and 60% of all metazoan protein coding genes are regulated by miRNAs in one or more cellular contexts[27, 28]. While miRNAs have been shown to be involved in most biological pathways or processes that are studied, they appear to be especially important in differentiation and in forming cell identity [29-31]. Consistent with this, many miRNAs appear to be expressed in distinct patterns in tissues in the metazoan body[32]. There are many examples of individual miRNAs that have strong impacts on development and phenotype, examples include the role of *lin-4* in nematode embryogenesis[1, 14, 33], the role of miR-430 in purging maternal transcripts from the zebrafish embryo[20] and even an example where a point mutation generates a miR-1 target site in the 3'UTR of the myostatin mRNA, causing a strain of especially muscular sheep[34].

Biogenesis

Drosha cleavage and nuclear export

Most miRNAs are transcribed by RNA polymerase II as long primary transcripts (pri-miRNAs) that are capped and polyadenylated and can be several kilobases in length[35, 36]. Each pri-miRNA contains one or more hairpin structures that are recognized and cleaved by the Microprocessor complex while the transcript is still in the nucleus[37] (se figure 1). This complex consists of the Drosha endonuclease and the DGCR8 dsRNA binding protein, which is necessary for recognizing the hairpin structure[38-41]. After the hairpin, also called the precursor miRNA (pre-miRNA), has been released from the pri-miRNA, it is exported to the cytosol by the Exportin5 nuclear export protein[42, 43].

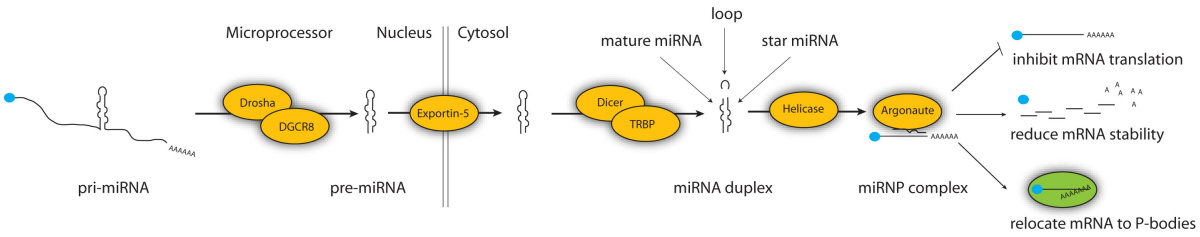


Figure 1: miRNA biogenesis.

Dicer cleavage

In the cytosol, the miRNA precursor hairpin is further recognized and cleaved by the endonuclease Dicer in complex with the TRBP dsRNA binding protein[44-48]. The characteristics of the hairpin before and after this cleavage is of importance for this thesis and will be described in detail here. Before the Dicer cleavage, the pre-miRNA hairpin is ~70 nucleotides (nts) long and consists of a terminal loop flanked by two arms that form a stem. The stem does not contain bifurcations, but typically 20% of the nucleotides in the stem are not base paired and form bulges (unpublished results). The entire hairpin is energetically stable compared with other non-coding RNAs of comparable length, like rRNAs and tRNAs[49]. After the Dicer cleavage, three products are released: The loop and the two strands of the stem[50]. The loop is typically of length 10-40 nts and is presumably rapidly degraded by exonuclease action. The two strands of the stem, both ~22 nts in length, remain bound to each other. Due to the endonuclease action, the two strands are offset thus that the duplex has 3' overhangs two nucleotides in length in both ends of the duplex.

Incorporation into the miRNP effector complex

The duplex in then unwound, and typically one of the strands is selectively bound to the Argonaute protein in the miRNP effector complex while the other strand is degraded. The strand that is less tightly base paired in the 5'end is more often incorporated into the effector complex[51, 52]. By definition, the strand that is more often incorporated is referred to as the 'mature' miRNA, while the strand that is more often degraded is the 'star' miRNA (sometimes these are referred to as the 'guide' and 'passenger' strands). In practice, the distinction between the mature and star strands is blurry. For instance, the ratios of incorporated mature vs. star strands can change during development in a given organism[53], and the ratios can change over evolutionary time, causing a reversal of the dominant strand[54]. Further, there is strong evidence that many miRNAs have mature and star sequences that are incorporated into the effector complex in comparable abundances and are both functional[53].

Alternative routes into the miRNA pathways

Many miRNAs are derived from the introns of protein coding genes and may be co-transcribed with host genes[55]. However, the expression of these miRNAs do not always correlate with the expression of the host genes[54], suggesting that the miRNAs are themselves post-transcriptionally regulated. Recent studies show that some short ~70 nts introns can undergo Dicer processing and enter the miRNA pathway without previous Drosha processing ('mirtrons'[56-58]). Further, we have recently shown that some snoRNAs with hairpin structures can enter the miRNA pathway and produce miRNA-like transcripts that can reduce expression of endogenous genes (see attached article[59]).

Target specificity

The miRNA 'seed'

Once the miRNA is incorporated into the miRNP effector complex, it can direct the complex to target sites in the 3'UTRs of mRNAs to degrade the mRNA or inhibit its translation. Unlike plant miRNAs which are thought to bind to mRNA target sites with almost full complementarity[60], animal miRNAs have only partial complementarity to their target sites. It was noticed early that especially the 5' end of miRNAs is important for the binding[61], in particular nucleotides 2-7 or 2-8 from the 5' end, sometimes referred to as the 'seed' or the 'nucleus'[62].

Computational target prediction

Although it is possible to identify likely miRNA targets by perturbing miRNA abundances in a model system and noting changes at the mRNA or protein level[29, 63-65], this has not until recently been performed for more than a few miRNAs. Thus the field of miRNA target identification has been dominated by computational prediction algorithms. The more accurate algorithms, like Pictar and TargetScan[27, 66, 67], identify miRNA target sites by searching 3'UTRs for occurrences of the sequence complementary to the seed of the miRNA. The confidence of the predictions is then increased by considering only sequence occurrences that are conserved in a number of species and by combining evidence from multiple occurrences in a given 3' UTR[27, 66, 67]. While for instance conservation scoring substantially improves the prediction accuracy, it also illustrates that we yet do not understand why some sequence occurrences are bound and others not (obviously the conservation information is not present in the given cell where the interactions take place). However, emerging technologies like CLIP-seq (see section on deep sequencing[68-71]) can

empirically identify binding sites for argonaute proteins and for RNA binding proteins that interact with the miRNP complex and may thus yield more of the missing pieces of the puzzle.

Mechanisms of miRNA-mediated repression

inhibition of translation

The early notion that miRNAs can have impact on protein abundances that cannot be explained purely by mRNA degradation has recently been validated by high-throughput proteomic studies[63, 64]. However, it is still hotly debated if this inhibition of translation is effected at the initiation of translation or at a post-initiation stage[72]. One model which favors inhibition at the initiation stage proposes that argonaute proteins inhibit binding of the eIF4E factor to the 5' cap, either by directly competing for binding or by recruiting other factors that compete[73]. Another model which favors the initiation stage proposes that the miRNP complex blocks association of the 60S subunit with the 40S pre-initiation complex[74]. In contrast, the models that favor inhibition at the post-initiation stage propose that the miRNP complex causes ribosomes to stall at or fall off the mRNA during elongation[75-77]. To complicate matters, these apparently contradictory models are all supported by evidence from various biochemical *in vitro* and *in vivo* studies in different model systems. A recent study has however shown that the GW182 protein, a core component of miRNP and of P-bodies (see below) contains three domains that are each sufficient to inhibit translation in tethering experiments[78]. Thus the possibility that these domains inhibit translation through different mechanisms may reconcile the apparently contradictory observations.

mRNA destabilization

It is well established that the miRNP complex can also destabilize mRNAs, causing their degradation[18, 20, 29]. A recent study indicates that inhibition of translation and deadenylation precedes the destabilization[79]. Since Argonaute, GW182 and de-capping enzymes are required for miRNA mediated destabilization[19], the next step is likely de-capping of the mRNA followed by degradation by exonuclease action. It is not yet clear what causes some miRNA-mRNA interactions to favor inhibition and others to favor destabilization, but it has been suggested that the number and positions of bulges in the miRNA-mRNA duplex play a role[80, 81].

localization to P-bodies

The P-bodies are cytoplasmic foci that are involved in mRNA degradation and storage[82-84]. It has been shown that argonaute proteins, miRNAs and their mRNA targets are enriched in P-bodies

and that there is a correlation between miRNA-mediated translational repression and the accumulation of mRNAs in the foci[85-88]. Depletion of proteins in the miRNA pathway causes a dispersal of the P-bodies[89, 90]. Reversely, depletion of proteins that form the scaffold of the P-bodies also cause dispersal of the foci, but has no effect on miRNA-mediated repression[89, 91]. Thus it appears that P-bodies are a possible effect rather than a cause of the miRNA function.

Functions

A description of miRNA biogenesis and mechanism of regulation does not confer what functions miRNAs have at the level of the cell, organism or evolution. Given that metazoans typically have hundreds of miRNA genes that together regulate 30-60% of all protein coding genes, and given the range of regulatory mechanisms available, it is difficult to make generalizations. However, a number of themes emerge from the literature, including:

miRNAs as switches

There are a number of examples where miRNAs work to purge cells of transcripts from earlier development programs, thus enforcing a clean switch from one developmental stage to the next. In *C. elegans* the heterochronic gene *lin-14* encodes a protein that is needed for the completion of the first larval stage (L1). However, unless the LIN-14 protein is depleted as the larva enters the second larval stage (L2), the first stage will be re-iterated[92]. The first miRNA to be described in any animal, *lin-4*, begins getting transcribed in the L1 to L2 transition and inhibits translation of the *lin-14* mRNA by binding to seven target sites in the 3'UTR[1, 14]. The switch function is clear: before the transition, *lin-4* miRNA is absent and the LIN-14 protein is present; after the transition the reverse is true. In zebrafish, miR-430 begins getting transcribed as the zygote transits from maternal to zygotic transcription. The miRNA accelerates the degradation of hundreds of maternal transcripts, thus delineating the transition[20]. This can be regarded as a switch function, since the effect of miR-430 is to reduce target expression to zero. Zebrafish Dicer mutants have several defects during gastrulation and brain morphogenesis. Interestingly, injection of mature miR-430 rescues these brain defects[93].

miRNAs as tuners

Recent high-throughput proteomic studies support the notion that many miRNA targets are only slightly down-regulated[63, 64]. This also holds for many target sites that are conserved, and therefore likely under positive selection. A possible explanation for this observation is that miRNAs can serve as an extra layer of post-transcriptional regulation, thus fine-tuning the output from the

transcriptional machinery. Mouse immunology can serve as a proof of principle that fine-tuning of protein output can have a strong phenotypic effect. In mouse lymphocytes, miR-150 modulates the expression of c-Myb, which promotes B cell survival[94]. Ectopic expression of miR-150 has subtle effects on the levels of c-Myb protein (30% reduction). This modest reduction, however, has a dramatic impact on the number of B cells in the mouse (more than four-fold reduction). However, miRNAs need not have single strong phenotypic effects to be important. For instance, when miR-1 (a miRNA expressed in muscles) is ectopically expressed in HeLa cells, the entire transcriptional profile changes to resemble that of muscle cells more[29]. Thus miRNAs can also impact the transcriptome through numerous 'soft' effects.

miRNAs as buffers

The switch function refers to miRNAs that either turn off gene expression or accelerate the turning off. The tuning function refers to miRNAs that reduce the mean gene expression. The buffering function refers to miRNAs that reduce the variance rather than the mean of gene expression. This function could theoretically help to make the expression of protein coding genes more robust and stable against stochastic fluctuations in transcription and translation efficiency and also against environmental influences. Such buffering could increase the connection between genotype and phenotype, and thus increase heritability[95-97]. The miRNA buffering function finds theoretical support from network models, in which many miRNAs are predicted to interact with transcription factors and target genes in regulatory networks that would stabilize gene expression (reviewed by Hornstein *et al.*[95]). However, there is yet little solid evidence to support that miRNAs functions as buffers (but see Wu *et al*.[96]). One problem is that laboratory experiments are designed to the minimize environmental influences that miRNAs should stabilize. Thus, experiments that simulate the stressful environment of nature might reveal more differences between say, wild-type animals and Dicer knockout animals. It might also be interesting to use real-time single cell imaging to investigate if Dicer knock-out cells have more individual variance in protein output than do control cells.

piRNAs:

In contrast with miRNAs which appear to be expressed in all animal tissues, piRNAs have only been detected in germline cells[98-106], in somatic cells of fly ovaries[107, 108] and in neoblast stem cells of planarian flatworms[109, 110]. piRNAs interact with Piwi proteins, a subgroup of the argonaute proteins which is essential for germline development and fertility[103, 111-116]. The primary function

of piRNAs is believed to be the silencing of transposable elements in the germline, although there is evidence that the piRNA pathway has also been adapted for other purposes, like post-transcriptional gene regulation[115]. The silencing is effected through slicing of transposon mRNA or through DNA methylation of the transposon genes[2, 117, 118]. Unlike miRNAs, the piRNA populations have high variation, consisting of millions of molecules of different sequence. These populations are mutable, and can change completely over a few days (mouse testes)[119] or between adjacent cells (fly ovaries)[107, 108, 117]. The first piRNAs were recently discovered (in 2006) and although piRNAs have now been detected in several species[21, 103-106, 109], they have only been systematically studied in mouse testes and in fly ovaries. In the following I will present the models that have been put forward to explain observations made in these two systems.

piRNAs in fly ovaries

piRNA populations have been investigated in fly follicle cells, which are somatic cells that support the fly germline cells, and in the fly oocytes.

piRNAs in fly follicle cells

Fly piRNAs are transcribed as long primary transcripts from hundreds of genomic clusters which typically have a high content of inactive transposons[117]. In the follicle cells, the primary transcripts are sliced into piRNA fragments and bound by the nuclear Piwi protein[107, 108]. These 'primary' piRNAs are characterized by being 24-30 nts in length and having a uracil in the 5'end. A large fraction of the piRNAs will have a sequence that is antisense to a transposable element. If the transcript of such an element if bound by sequence complementarity to the piRNA, the transcript will be sliced by the Piwi protein. The *flamenco* piRNA cluster, which is necessary for silencing *gypsy* elements[120], is highly expressed in the follicle cells. Interestingly, *gypsy* retrotransposons have their reproductive cycle in the follicle cells and then make virus particles to infect the germ line[121-123]. This is possibly a way for the *gypsy* elements to avoid the more elaborate defenses of the germ line cells (see below), and the *flamenco* cluster expression in the follicle cells may be an evolutionary response to this.

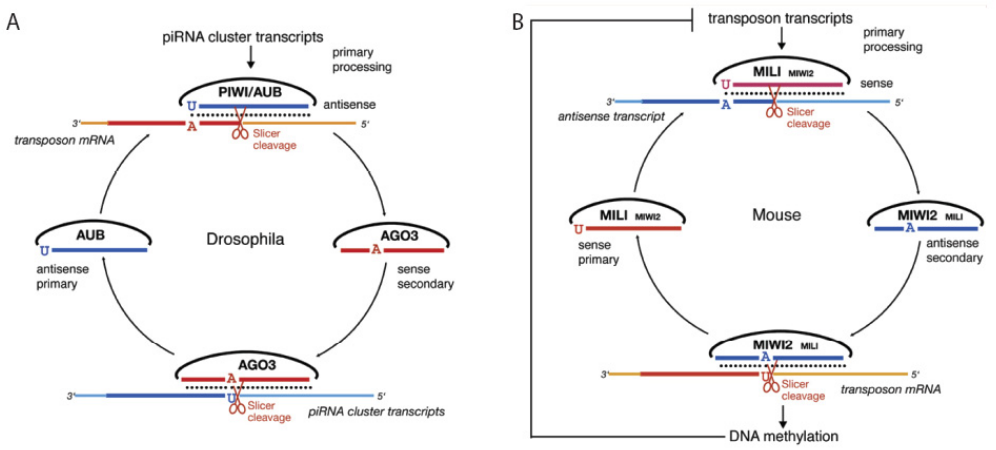


Figure 2. The 'ping-pong' model of piRNA biogenesis and slicing of transposons in A) fly oocytes and B) mouse pre-pachytene testes. From Aravin *et al.*[119]

piRNAs in fly oocytes

In the fly oocytes, Piwi performs a similar nuclear function as in the follicle cells. In addition two more proteins from the Piwi family, Aubergine and Ago3, are located in the cytoplasm, primarily in the 'nuage' perinuclear foci[117]. Aubergine, like Piwi, slices piRNA fragments from the piRNA cluster transcripts. When an Aubergine interacting piRNA binds to a transposon transcript it will slice it, and a fragment of the transposon transcript will be loaded to Ago3. When this Ago3 interacting piRNA binds a cluster transcript it will slice it, and a fragment of the cluster transcript will be loaded to Aubergine (see figure 2A). In this manner, transposon transcripts and cluster transcripts are iteratively sliced and loaded into the two Piwi family members, causing a depletion of both types of transcripts[117, 118]. This model is referred to as the 'ping-pong' amplification loop and is supported by two observations. First, piRNAs that associate with Aubergine and Ago3 tend to overlap in the 5' ends by exactly ten nucleotides, as would be expected if they were iteratively generated by activity of a slicer protein domain[117, 118]. Second, while piRNAs that bind to Aubergine typically has a uracil in the 5' end, the piRNAs that bind to Ago3 typically has an adenosine at position ten, as expected if the piRNAs were defined by being base-paired with the Aubergine piRNAs by ten nucleotides. In flies, the initiating or 'primary' piRNAs, bound to Aubergine, are sliced from the piRNA cluster transcripts and are therefore typically antisense in sequence to transposons. The 'secondary' piRNAs, bound to Ago3, are sliced from transposon transcripts and therefore contain transposon sense sequences. The piRNA pathway constitutes a transposon defense with both static and adaptable components. The defense is static since the amplification loop is only initiated if a transposon with sequence complementary to a piRNA cluster is encountered. On the other hand the defense is adaptable since active transposons can insert into clusters and thus enter the 'memory' of the pathway[117]. This is consistent with

observations that fly transposon defense is mounted after a variable number of novel transposon copies have inserted into the genome[120].

piRNAs in mouse testes

Mouse gametes loose most of their methylation patterns after fertilization[124]. Around seven days after fertilization spermatogenesis begins in the mouse male embryo as the primordial germ cells migrate into the gonadal compartments and expand through mitotic division[125]. This division stops around 15 days after fertilization as DNA methylation of transposable elements and imprinted loci is re-established (see figure 3). Mitotic division resumes three days after birth and around ten days after birth the meiotic divisions begin that will produce the mature sperm. The meiosis can be divided into (1) the leptotene in which duplicated chromosomes condense, (2) the zygotene in which extensive pairing and formation of synaptonemal complexes occur, (3) the pachytene in which crossing over occurs, (4) the diplotene in which homologs begin to separate and (5) the diakinesis in which the chromosomes move apart. In mouse embryonic testes, piRNA biology has been investigated just before birth as DNA re-methylation occurs (pre-natal), ten days after birth as meiosis initiates (post-natal pre-pachytene) and in adult mice (pachytene).

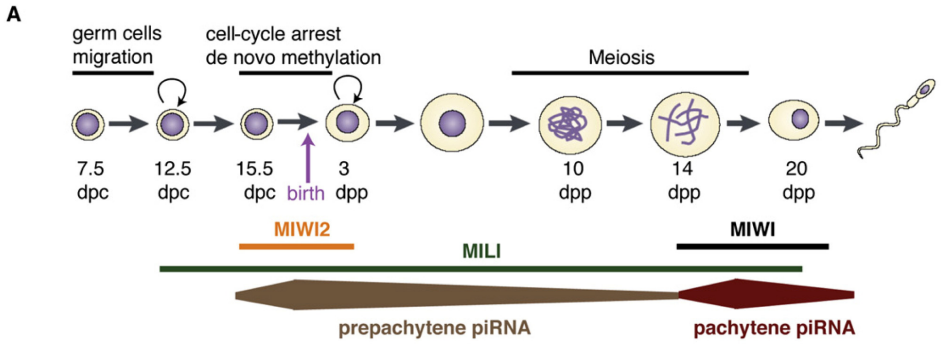


Figure 3. Mouse spermatogenesis and Piwi protein expression. dpc, days after fertilization, ddp, days after birth. From Aravin *et al.*[119]

Pre-natal piRNAs

In the mouse embryonic testes just before birth, the cytoplasmic Piwi family MILI protein cleaves primary piRNA fragments from transposon mRNAs (in particular LINE-1 and IAP elements which are particularly active at this time point)[119]. These piRNAs then bind to piRNA cluster transcripts that are anti-sense to transposon sequence, causing them to be sliced and loaded into the mostly nuclear Piwi family MIWI2 protein. As the MIWI2 interacting cluster fragments then

bind to transposon transcripts, a ping-pong amplification cycle initiates[119]. However, the cycle is reversed compared to fly, as the initiating primary piRNAs are sliced from transposon transcripts and the secondary piRNAs are sliced from cluster transcripts (see figure 2B). This is also reflected in piRNA sequence characteristics as the mouse primary piRNAs have a beginning uracil and are sense to transposon sequence while the secondary piRNAs have an adenosine at position ten and are antisense to transposon sequence. It is believed that the interaction between the cytoplasmic MILI and the nuclear MIWI2 takes place in perinuclear structures that resemble the fly nuage. MIWI2 then shuttles to the nucleus and initiates transposon re-methylation[119]. This is supported by observations that the germline cells of MILI or MIWI2 mutants display loss of LINE-1 and IAP transposon methylation, have increased transcript levels of these transposons, show evidence of double-stranded DNA breaks indicative of novel transposon insertions, and arrest during meiosis[2, 111, 116, 126]. These mutants cannot complete spermatogenesis and are sterile[111, 116]. It is not known how MIWI2 initiates the re-methylation, but there is evidence that the protein does not directly interact with DNA methyltransferases[119].

Post-natal pre-pachytene piRNAs

After the genome has been re-methylated and as meiosis initiates in the germ line, MILI is the only Piwi family protein that is expressed in the testes of the mouse pup[119]. MILI continues the ping-pong cycle alone, as evidenced from sequence characteristics and the typical ten nucleotide overlap of piRNAs. The piRNA clusters that are expressed at this stage are different from those just before birth, meaning that the piRNA populations are also almost completely distinct[2, 119]. Also, fewer LINE-1 and IAP derived piRNAs are observed, while more piRNAs from SINEs and from mRNA exons are present. This probably reflects changes in abundances of transcripts that are available for slicing. Since primary mouse piRNAs are directly sliced from transposon mRNA, the piRNA clusters appear to be less important than in fly, where cluster piRNAs initiate the defense[117]. This reversal of the amplification loop may reflect a more flexible response to expansion of transposable elements in mammals. On the other hand, it pushes forward the question how the MILI protein can distinguish mRNAs from transposable elements from those of the host. Given that many piRNAs are derived from mRNAs of host protein coding genes, it appears that the MILI does in fact sample the transcriptome to some degree[119].

Pachytene piRNAs

In the adult testes, MILI is still expressed as is now the MIWI protein[115, 119]. Again the piRNA clusters are different from those in the mouse pup. The piRNAs no longer show evidence of ping-pong amplification and are strongly depleted in transposon sequence, suggesting that the primary role of piRNAs in the adult testes is not transposon silencing[99]. Consistent with this, the MIWI

protein is necessary for germ line maintenance and fertility, but there is no evidence that the protein is involved in transposon silencing[115, 127]. However, MIWI has been shown to associate with piRNAs and mRNAs in RNPs and in polysomes[128]. Further MIWI is necessary for the expression of some mRNA transcripts, suggesting that it is involved in post-transcriptional regulation of host genes[115]. However, this regulation has not been studied further.

Nematode 21U-RNAs

In the *C. elegans*, the Piwi homologs Prg-1 and Prg-2 bind to small RNAs that have a beginning uracil and that are 21 nts in length. These 21U-RNAs are expressed in the nematode germ line and are necessary for fertility[129-131]. While this constitutes strong evidence that the 21U-RNAs are the piRNAs of nematodes, the biogenesis appear to be quite different. The 21U-RNAs are described from ~20.000 distinct loci that each have an characteristic upstream motif, suggesting that they are individually transcribed[50]. They also do not display the ten nts overlap that is indicative of ping-pong amplification. Further, while there is evidence that some 21U-RNAs silence Tc3 transposons[130, 131], the function of the majority of these small RNAs remain unknown. Like the mouse pachytene piRNAs they are as a population depleted in transposon sequence, suggesting that 21U-RNAs might perform functions other than transposon silence, such as mRNA regulation.

piRNA summary

The models of piRNA biology in mouse testes and fly ovaries demonstrate that piRNA populations and their functions can change dramatically over a few days or even between adjacent cells. It will be interesting to see if more systematic studies of piRNAs will bring forth unifying themes or if the complexity of piRNA biology will increase linearly with the amount of data produced.

Deep sequencing:

Deep sequencing is an emerging high-throughput technology (the first deep sequencing machines became commercially available in 2005). The technology allows sequencing of DNA, cDNA or small RNAs ligated to DNA adapters. Common to the available deep sequencing platforms is miniaturization and massive parallelization which allows for the simultaneous sequencing of millions of DNA or RNA molecules[132]. This means that one machine can sequence literally billions of nucleotides in a few days, which represents several orders of magnitude improvement over the previous generation of capillary Sanger sequencers. Deep sequencing has numerous applications, which include but are not limited to:

-*genome re-sequencing*. This is a term for the re-sequencing of genomes that have previously been sequenced and assembled. This is a useful applications for population genetics, studies linking genotypes to phenotypes and potentially even for personalized medicine. The Life Sciences / Roche company recently sequenced the genome of Dr. James Watson for less than 1.5 million dollars using only the 454 deep sequencing platform[133].

-*de novo genome sequencing.* Novel genomes can be deep sequenced, but given that the sequencing reads produced by deep sequencing are shorter than those produced by Sanger sequencing (table 1), assembly remains a challenge. None the less, the oil palm tree genome was recently sequenced and assembled using only the 454 deep sequencing platform (not yet published).

-*genome bisulfite sequencing.* This is a method to selectively sequence the parts of the genome that are DNA methylated. It is one way in which deep sequencing can survey epigenetic information (reviewed in Pomraning *et al.*[134])

-*RNA-seq*. Deep sequencing of mRNAs generates several levels of information. First, the number of times a mRNA is sequenced correlates well with transcript abundances as estimated from qPCR[135]. Compared with arrays, this 'digital gene expression' is unbiased since it does not depend on pre-spotted probes on an array. Second, when exon-exon junctions are sequenced, information on splice variants is also yielded. Third, sequence information such as SNPs or RNA editing can also be obtained (reviewed in Wang *et al.*[136]).

-*CHIP-seq.* This method uses immunoprecipitation to pull-down transcription factors and sequence the DNA that they bind to (reviewed in Park[137]).

-*CLIP-seq.* Similar to CHIP-seq, but the method uses pull-down of RNA binding proteins cross-linked to RNA[68-71].

-*small RNA sequencing.* Deep sequencing allows for the sequencing of millions of small RNAs in a sample. This has made it possible to discover small RNAs that were previously below detection limits and to do high-throughput small RNA digital gene expression. It has also made possible the study of small RNA populations that have a high degree of sequence diversity, like the piRNAs.

Currently three deep sequencing platforms are commercially available and wide used: the 454 / Life Sciences platform, the Solexa / Illumina platform and the ABI SOLiD platform. The platforms differ in the technology used as well as in the performance statistics:

	454 FLX titanium	Illumina GA IIx	ABI SOLiD 3
Total output	0.5 billion nts	5 billion nts	15 billion nts
Read length	400 nts	35 nts	50 nts
Number of reads	>1 million	150 million	300 million
Time to prepare library	2-3 days	2-3 days	2 weeks
Time per sequencing run	10 hours	2 days	7 days
Cost per sequencing run	10.000 €	5.000 €	5.000 €

Table 1. Summary of deep sequencing platforms. The statistics shown here are for small RNA sequencing – the statistics may vary slightly when the platforms are used for other applications.

454 / Life Sciences

technology: First adapters are ligated to the fragmented DNA or cDNA, or to the small RNAs. Then the ligation products are bound to micrometer beads under conditions that favor the binding of one product per bead. The beads are captured in droplets of oil that contain enzymes for emulsion PCR reaction. Inside every droplet a PCR reaction occurs, resulting in each bead being covered by millions of identical copies of the captured ligation product. Subsequently the beads are deposited into 1.6 million micrometer wells on a fibreoptic slide, one bead per well. The slide is mounted in a flow chamber through which sequencing reagents flow, and pyrosequencing takes place in each well. Every time a given nucleotide is incorporated, light of a given wave length is emitted. The light emissions are detected and translated into nucleotide sequences (www.454.com and Shendure and Ji[132]).

pros: The main advantage of the 454 platform is that long (~400 nts) DNA or cDNA molecules can be sequenced. This makes any kind of sequence assembly much easier. This may be especially important for studies on transcript splice variants and genome *de novo* sequencing.

cons: The disadvantages of the 454 platform include a) the number of molecules sequenced in a single run is relatively low (~1 million) b) the number of nucleotides sequenced in comparably low (~0.5 billion nts) c) the platform has difficulties in determining the number of nucleotides in homonucleotide stretches (e.g. poly-As). Further, the light signal emitted from such stretches can spill over into nearby wells, causing further sequencing errors (personal communication, Azra Krek). Last, the reagents for a 454 sequencing run are approximately twice as costly as those consumed by the other two platforms (table 1).

Solexa / Illumina

technology: First adapters are ligated to the fragmented DNA or cDNA, or to small RNAs. The ligation products are attached to the surface of a flow cell, to which PCR enzymes and nucleotides are added. Aside from the ligation products the flow cell is also covered by a dense lawn of primers that are complementary in sequence to the adapters. The adapters will bind to these, making each ligation products form a bridge over which amplification occurs. After numerous rounds of bridge amplification, the flow cell will be covered by millions of clusters that each contains thousands of copies of one ligation product. Last, enzymes and fluorescently labeled nucleotides are added and sequencing by synthesis takes place in each cluster on the flow cell. A laser excites the nucleotides that in incorporated in each cycle in each cluster, and the light emissions are translated into nucleotide sequences (www.illumina.com and Shendure and Li[132]).

pros: A single flow cell produces ~150 million sequencing reads of short length (table 1), making it an ideal platform for small RNA sequencing. Consistent with this, most (67/97) small RNA studies with dataset depositions at the GEO (Gene Expression Omnibus) database have been undertaken using the Solexa platform.

cons: The rate of sequencing errors increase towards the end of the deep sequencing reads. This can in some cases make it difficult to map the reads full-length to the genome or to computationally remove the 3' adapters from deep sequenced small RNAs.

ABI SOLiD

technology: Adapters are ligated to the fragmented DNA or cDNA, or to small RNAs. Similar to the 454 platform, the ligation products are bound to beads and emulsion PCR reaction takes place in microreactors, such that each bead gets covered in millions of copies of the same ligation product. The resulting copies are then covalently bound to a glass slide, such that identical copies from one bead locate to one cluster on the slide. Then primers complementary to the adapter sequence are added and extended with di-base probes that compete for ligation to the primer. The di-base probes are fluorescently labeled and indicate the sequence of di-nucleotides of each cluster of identical ligation products on the glass slide. After the di-base probes have extended the primer to the length of the ligation product template, the primer and the probes are removed and a new primer, offset by one nucleotide, is added along with di-base probes. This is repeated five times, such that each nucleotide of a given ligation product is interrogated by multiple di-base probes (www.appliedbiosystems.com and Shendure and Li [132]).

pros: A single SOLiD run can produce more than 300 million reads. Further, the di-base color encoding makes it easier to discern sequencing errors from nucleotides that differ from the reference sequence due to SNPs or RNA editing. This may make the platform particularly suitable for genome re-sequencing, where identification of SNPs may be particularly important.

cons: Importantly, it is time-consuming to prepare samples for SOLiD sequencing. Further, the di-base color encoding makes it necessary to have dedicated computational tools for most downstream analysis.

Analyzing small RNA deep sequencing data

miRNAs

The legacy of conventional cloning and Sanger sequencing

Since the first systematic studies to identify new miRNAs, sequencing has been the method of choice for miRNA discovery[24-26]. Before deep sequencing, researchers would using conventional cloning and Sanger sequencing for this purpose. The resulting sequences were then mapped back to the reference genome to identify which loci they were transcribed from. It was soon recognized that many sequence mapped to rRNA, tRNA and protein coding genes and were likely short RNA[ii] degradation products. To identify the miRNA fraction, researchers discarded all sequences that mapped to these annotations. Further, if a given sequence is a sequenced mature miRNA, then it will be expected that the sequence has been cleaved out of either arm of a miRNA precursor hairpin. Therefore, the researchers required that the flanking genomic sequence of a miRNA candidate should be predicted to form a hairpin when folded with an RNA structure prediction algorithm. If the candidate miRNA star sequence was also detected, it was seen as confounding evidence[24-26].

Early miRNA deep sequencing analysis

Deep sequencing of small RNAs and the downstream analysis of these is a direct continuation of this tradition. The extra sequencing depth opens up new possibilities, like the identification of very lowly expressed miRNAs. At the same time, it gives new challenges. For instance, while a set of Sanger sequenced small RNAs might map to thousands of genomic loci that each need to be analyzed, a set of deep sequenced small RNAs typically map to millions of loci. Further, even the deep sequenced RNAs that map to miRNA loci do not all correspond to canonical Drosha/Dicer products (see Discussion section). These challenges demand more sophisticated algorithms. When I in the beginning of my Ph.D. first set out to identify novel miRNAs in deep sequencing data, there were no publicly available algorithms for this purpose. In fact, there were even no tools specialized for mapping short RNAs against a reference genome, meaning that even this initial step was non-trivial (such tools however emerged, e.g. Berninger et al.[7]). There were a few studies reporting novel miRNAs from deep sequencing data. One of these studies employed elaborate comparative genomics to identify conserved hairpin structures[138]. One other study used a rules-based approach but was not described in enough detail that it could be implemented from the manuscript[50]. These two algorithms were not publicly available.

[ii] The term 'short RNAs' is here denotes all cellular transcripts of length <40 nts. This includes regulatory small RNAs and degradation products of longer transcripts like mRNAs, rRNAs, tRNAs etc.

Our contribution

Supervised by Nikolaus Rajewsky I developed miRDeep, an algorithm for identifying known and novel miRNAs in deep sequencing data. Using this algorithm we have discovered hundreds of novel miRNA genes in a dozen species[110, 139, 140] (and unpublished results). The algorithm is publicly available and has also been used by several research groups to identify more novel miRNAs[141-145].

Later miRNA deep sequencing analysis

Since miRDeep was published, two new algorithms for identifying miRNAs in deep sequencing data have become publicly available. mirCat[146] is specialized for plant data, while miRanalyzer[147] is specialized for animal data. miRanalyzer recovers known miRNAs in unseen data with a sensitivity comparable to miRDeep, but has a false positive rate that is approximately two orders of magnitude higher than miRDeep[110, 147]. There also exists a number of miRNA gene finding tools that are not specialized for deep sequencing data (reviewed in Mendes *et al.*[148])

piRNAs

Because piRNA populations consist of vast numbers of distinct sequences, the systematic study of them was not possible before deep sequencing became available. When I first analyzed planarian piRNA data, there were a number of studies on the subject which all carefully described the computational analysis performed (e.g. [2, 106, 117]). This analysis includes mapping sequenced piRNAs to the genome, identifying the ones that can be mapped to a single genomic locus, and using these unambiguous mappers to identify the piRNA clusters. The analysis also included an investigation of sequence and length biases in the subpopulations of piRNAs. However, the piRNAs had in these studies all been isolated by immunoprecipitation or other purification of different Piwi proteins, meaning that they were piRNAs by definition (piwi interacting RNAs). The planarian small RNA data I analyzed was from total RNA, meaning that the initial challenge was to computationally isolate the piRNA fraction. I also made a partly successful attempt to further divide the piRNA fraction into subpopulations using computation. There does not exist any publicly available algorithm for piRNA analysis.

PUBLICATIONS

Friedländer *et al.*, 'Discovering microRNAs from deep sequencing data using miRDeep', Nature Biotechnology (2008)

Friedländer *et al.*, 'High-resolution profiling and discovery of planarian small RNAs', PNAS (2009)

Ender *et al.*, 'A Human snoRNA with MicroRNA-Like Functions', Molecular Cell, (2008)

Stoeckius *et al.*, 'Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression', Nature Methods, (2009).

Discovering microRNAs from deep sequencing data using miRDeep

Marc R Friedländer[1], Wei Chen[2], Catherine Adamidi[1], Jonas Maaskola[1], Ralf Einspanier[3], Signe Knespel[1] & Nikolaus Rajewsky[1]

The capacity of highly parallel sequencing technologies to detect small RNAs at unprecedented depth suggests their value in systematically identifying microRNAs (miRNAs). However, the identification of miRNAs from the large pool of sequenced transcripts from a single deep sequencing run remains a major challenge. Here, we present an algorithm, miRDeep, which uses a probabilistic model of miRNA biogenesis to score compatibility of the position and frequency of sequenced RNA with the secondary structure of the miRNA precursor. We demonstrate its accuracy and robustness using published *Caenorhabditis elegans* data and data we generated by deep sequencing human and dog RNAs. miRDeep reports altogether ~230 previously unannotated miRNAs, of which four novel *C. elegans* miRNAs are validated by northern blot analysis.

Animal genomes harbor numerous small, noncoding miRNA genes believed to post-transcriptionally regulate many protein-coding genes to influence processes ranging from metabolism, development and regulation of the nervous and immune systems to the onset of cancer[1]. Despite concerted efforts to discover and profile miRNAs, even the number of miRNAs in the human genome remains controversial, with estimates ranging from a few hundred[2] to tens of thousands[3]. Traditional experimental approaches to miRNA discovery have relied on cloning and Sanger sequencing protocols[4] and human and murine miRNAs have been profiled in hundreds of cDNA libraries from dozens of tissues[5].

However, the vast dynamic range of miRNA expression (from tens of thousands to a few molecules per cell) complicates profiling of miRNAs expressed in low numbers. A complementary approach, involving miRNA discovery by computational predictions that analyze genomic DNA for structures that resemble known miRNA precursors[6], is compromised by sensitivity problems and substantial numbers of false positives[6]. Therefore, purely computational approaches require experimental follow-ups, which are again difficult for miRNAs with low expression levels in the sample.

'Deep-sequencing' technologies have opened the door to detecting and profiling known and novel miRNAs at unprecedented sensitivity. Next generation sequencing platforms, such as those from Solexa/Illumina and 454 Life Sciences/Roche, can sequence DNA orders of magnitude faster and at lower cost than Sanger sequencing and are evolving so rapidly that increases in sequencing speed by at least another order of magnitude seem likely over the next few years. Although the Solexa/Illumina system can produce ~32 million sequencing reads in one run, read length is currently limited to 35 bp. In contrast, the current 454 platform yields reads up to 200 bases each, although the number of reads is an order of magnitude less than that of Solexa/Illumina. The nature of sequencing errors also contributes further to the different output characteristics of the two approaches.

Despite the ability of both technologies to sequence—and thus to detect—miRNAs at previously unmatched throughput, deep sequencing presents formidable computational challenges and suffers from biases such as those arising from the preparation of small RNA libraries. Even mapping deep-sequencing reads to the genome is itself not trivial, as no animal genome besides that of *C. elegans*, has been sequenced completely. Moreover, sequencing errors and polymorphisms, as well as RNA editing and splicing are but some of the factors that contribute to ambiguity. Although currently almost all of these problems remain mostly unsolved, deep sequencing can successfully survey the small RNA contents of animal genomes with unmatched sensitivity[7–15].

When profiling small RNAs with deep-sequencing technology, separating miRNAs from the pool of other sequenced small RNAs or degradation products is a central problem that is often not described or only partially addressed[8,9]. Furthermore, despite a growing need to analyze deep-sequencing data, there is no publicly available algorithm to detect miRNAs in these data.

miRDeep, our publicly available software package, can be used to solve this problem at least in part. Importantly, it also includes stringent statistical controls to estimate the false positive rate and the sensitivity of miRDeep predictions. Therefore, users can not only run miRDeep on their own deep-sequencing data to detect known and novel miRNAs, but can also estimate the quality of their results. At the heart of miRDeep is the idea of detecting miRNAs by analyzing how sequenced RNAs are compatible with how miRNA precursors are processed in the cell. As deep sequencing permits statistical analysis of this model, one can assign a score of the likelihood that a detected RNA is indeed a mature miRNA. Therefore, the foreseeable advances in sequencing capacity of deep-sequencing technologies should further boost the power of miRDeep. In order to address an ongoing discussion about the importance of nonconserved miRNAs[16] and to be as unbiased as possible, we designed miRDeep to detect miRNAs without cross-species comparisons. Finally, given the rapid evolution of deep-sequencing technology,

[1]Max Delbrück Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, D-13125 Berlin-Buch, Germany. [2]Department of Human Molecular Genetics, Max Planck Institute for Molecular Genetics, Ihnestrasse 73, D-14195 Berlin, Germany. [3]Institute of Veterinary Biochemistry, Freie Universität Berlin, Oertzenweg 19b, D-14163 Berlin, Germany. Correspondence should be addressed to N.R. (rajewsky@mdc-berlin.de).

Published online 7 April 2008; doi:10.1038/nbt1394

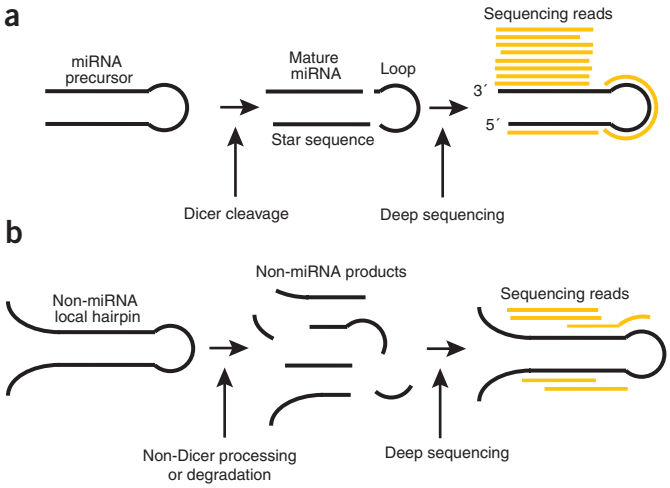


Figure 1 Analyzing the compatibility of sequenced RNAs with miRNA biogenesis. (**a**) Each of the RNA products generated after a stable miRNA precursor is cleaved by Dicer—the mature miRNA sequence, the star sequence and the loop[2]—has a certain probability of being sequenced. When miRDeep maps the sequenced RNAs ('reads') to the genome and to the corresponding predicted miRNA precursor hairpin structure, read sequences map to the positions reminiscent of the three Dicer products. However, the mature sequences are generally more abundant in the cell and are therefore also sequenced more frequently than the loop and star sequence RNAs. Thus, the statistics of the read positions and frequencies of the reads within the stable hairpin (the 'signature') are highly characteristic for miRNAs and are scored by miRDeep. The power of miRNA discovery by miRDeep is proportional to the depth of sequencing. (**b**) Large numbers of hairpins that are not processed by Dicer are also transcribed from metazoan genomes. These hairpins can also produce short RNAs, either through non-Dicer processing or through degradation. However, when the reads that originate from such sources are mapped back to the secondary structure, they will likely map in a manner that is inconsistent with Dicer processing.

we designed miRDeep to be as flexible as possible and tested it using both Solexa- and 454-derived data from human, the domestic dog and *C.elegans*—animals from the two main branches of Bilateria, representing very different genomic complexity.

RESULTS

miRDeep scores according to a model of miRNA biogenesis

Metazoan miRNA genes are transcribed either as single genes, or in clusters, or intronically as part of protein-coding transcripts[2]. Hairpins within the primary miRNA gene transcript are typically, but not always, recognized and cut by the endonuclease Drosha in the cell nucleus to produce miRNA precursors. These are then exported to the cytosol, where the hairpin structure is cut by the endonuclease Dicer at relatively fixed positions[17–19]. The hairpin processing by Dicer releases three products of largely invariant lengths (**Fig. 1**). One of these is the loop of the hairpin, which is degraded as a by-product. The two other products form a duplex, which is subsequently unwound by helicase activity. One of the strands in the duplex, the so-called star strand, is typically degraded, whereas the mature miRNA strand is taken up into the microribonucleoprotein complex (miRNP)[19]. The mature miRNA sequence functions by guiding miRNP to target mRNAs by partial sequence complementarity. The approximately six nucleotides starting at position two from the 5′ end of the mature sequence are particularly important for target recognition[20]. miRNP regulates the mRNA transcript by inhibiting translation or decreasing its stability[19].

An overview of the miRDeep algorithm is shown in **Figure 2**. Briefly, after the sequencing reads are aligned to the genome, the algorithm excises genomic DNA bracketing these alignments and computes their secondary RNA structure. Plausible miRNA precursor sequences are then identified and, in the core part of the miRDeep algorithm, scored for their likelihood to be real miRNA precursors. The output is therefore a scored list of known and novel miRNA precursors and mature miRNAs in the deep-sequencing sample, as well as estimates for the number of false positives.

In more detail, miRDeep initially investigates the secondary structure of each potential precursor as well as the positions of the reads that align to it. Next, a filtering step discards potential precursors that are grossly inconsistent with miRNA biogenesis. For the remaining (typically thousands of) potential precursors, miRDeep then probabilistically integrates deep-sequencing information based on a simple model for miRNA precursor processing by Dicer (**Fig. 1a**,**b**). If a sequence is an actual miRNA precursor that is expressed in the deep-sequencing sample, then one expects that one or more deep-sequencing reads correspond to one or more of the three products—the mature miRNA sequence, the star sequence and the loop (**Fig. 1a**)—released when the precursor is cut by Dicer[8]. Further, it is expected that only very few, if any, reads do not correspond to these three products. Reads originating from miRNA Dicer products have relatively invariant lengths and relative positions, and therefore high information contents. If an miRNA precursor candidate is part of an actual transcript, but not a Dicer substrate, then deep-sequencing reads will not fit into this model of processing. Often, the reads will originate from staggered degradation products of stochastic lengths and positions (**Fig. 1b**).

The miRDeep core algorithm scores each potential miRNA precursor for the combined compatibility of energetic stability, positions and frequencies of reads with Dicer processing. A number of features contribute to the score. In general, the greater the number of deep-sequencing reads corresponding to the mature or star products, the more likely the sequence is to be an miRNA precursor. The presence of one or more reads corresponding to the star sequence, taking into account the short 3′ duplex overhangs characteristic of Drosha/Dicer processing, adds to the score separately. As miRNA precursors are more stable than nonprecursor hairpins[21], both the relative and absolute stabilities of the structure also contribute to the score. Finally, the 5′ ends of mature miRNAs are often conserved across vast phylogenetic distances[22,23]. If the 5′ end of the potential mature sequence is identical to that of a known mature sequence, the score can optionally be increased. The probabilities of all features contributing to the score are estimated by parameter fitting to known and background miRNA precursors. These parameter fits were stable when separately analyzing data sets from animals spanning large phylogenetic distances, strongly suggesting that miRDeep does not overfit. In sum, the algorithm assigns each sequence a log-odds score, which indicates the probability that the sequence is a true miRNA precursor instead of a background hairpin. In what follows, we refer to the number and relative position of reads in a potential miRNA precursor as the 'signature'.

Statistical evaluation of miRDeep results

As many genomes contain large numbers of sequences that could fold into hairpin structures if transcribed (for instance, the human genome contains at least 11 million hairpins[6]) and most deep-sequencing reads originate from loci that are not miRNA genes (unpublished results), any algorithm that predicts miRNAs by intersecting deep sequencing data with secondary structure information risks producing vast numbers of false positives. We thus employed several stringent controls to estimate the sensitivity and the number of false positives per genome-wide analysis.

We estimated the sensitivity as the fraction of known mature miRNA sequences (from miRBase version 10.0 (ref. 24)) represented by at least one read in the raw deep-sequencing data sets recovered in the final predictions. Simple sequence matching is used to find known miRNAs in the data sets. As sequencing reads representing miRNA sequences often have untemplated nucleotides in the 3′ end[8,25], mismatches in the last three nucleotides are tolerated.

miRDeep scores each potential precursor by analyzing its read signature and its structure. We estimated the false-positive rate by running miRDeep on our input set of structures and signatures as usual, except that we randomly permuted the signature and structure pairings in the input data set. For example, if a read in a potential miRNA precursor A resides at relative position five (from the 5′ end), then it will be assigned to another potential miRNA precursor B, also at position five. All reads in A will be mapped to B in this manner. This control precisely tests our model hypothesis that for true miRNAs, the structure (the hairpin) is recognized by Dicer and therefore causes the signature. By permuting the structure and signature pairings, we thus simulate the null hypothesis that the two are independent. Analysis of multiple independent permutation runs furthermore yields the s.d. of the estimated mean number of false positives.

Our test is conservative in that it tends to overestimate the number of false positives. Many of the actual miRNA precursors have a large number of reads that map consistently with our model of miRNA processing by Dicer. When the signatures of these precursors are combined with unstable background hairpins, the large score contribution of the signature causes the overall score to exceed the cut-off. In other words, a significant fraction of the estimated false positives are caused by actual miRNA signatures through a 'hitchhiking effect'. Therefore, our false-positive estimates are likely an upper limit to the true number of false positives.

miRDeep handles heterogeneous input data robustly

Deep-sequencing data sets are very heterogeneous. Different genomes have different transcription profiles and long transcripts may be sequenced at the ends only, or represented by sequences of their degradation products. Some genomes transcribe short functional noncoding transcripts, such as endogenous small interfering RNAs or repeat-associated interfering RNAs[12,26]. Owing to their similar lengths, these can be particularly difficult to distinguish from miRNAs. Moreover, bias can be introduced during sample preparation where small RNAs are isolated and ligated with specific adapters. Finally, sequencing technologies vary in the frequency and types of sequencing errors, in the maximum length of the sequence reads and in the number of reads produced.

We have implemented miRDeep in a flexible, probabilistic manner such that miRNA precursors with single noncharacteristic features can be recovered if they display other characteristics. Besides testing the ability of miRDeep to detect known and novel miRNAs, we also wanted to assess how robustly miRDeep handles heterogeneous data. We therefore obtained *C. elegans* deep-sequencing data from the GEO database, and produced two more data sets ourselves by deep sequencing a dog lymphocyte sample and a human cell line. Together, these data sets represent Protostomes and Deuterostomes with very different genome sizes and transcriptional profiles. Further, the data sets were produced by different laboratories, using 454 sequencing or Solexa sequencing. The core miRDeep algorithm was run on the three data sets with identical parameter settings, except for the score cut-off parameter.

miRDeep detects novel miRNAs in previously mined data

The relatively small (~100 Mb) genome of *C. elegans*—the organism in which miRNAs were first discovered[27,28]—has been intensively mined for miRNA genes using both computational and experimental

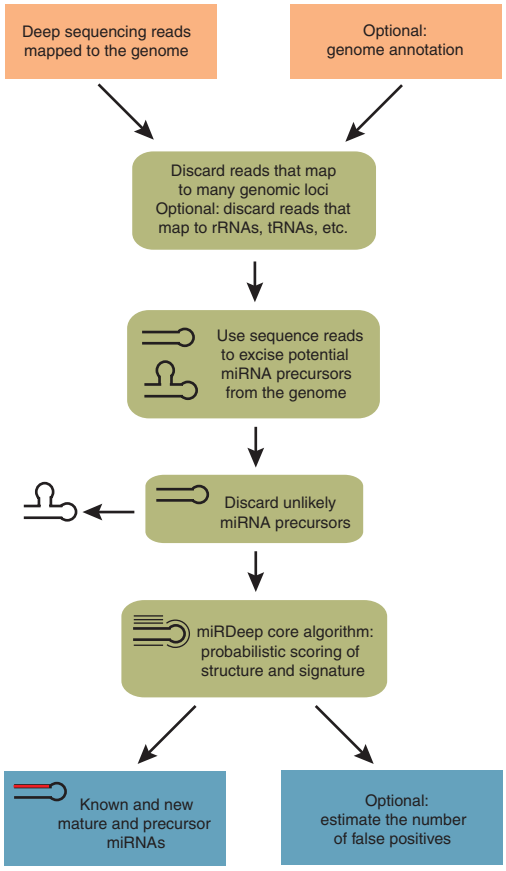


Figure 2 Flowchart diagram representing the miRDeep software package.

methods[29,30]. Specific detection of miRNAs in *C. elegans* is difficult, as the transcriptome has a large fraction of small RNAs, such as endogenous small interfering RNAs and 21U-RNAs[8] that can potentially cause many false positives. Our first data set comprised pooled reads from several 454 sequencing runs on *C. elegans* mixed-population small RNA samples[8,12], obtained from the GEO database.

The deep-sequencing reads were aligned to the *C. elegans* genome. Reads that aligned to more than five genomic positions, or to University of California Santa Cruz (UCSC) annotations of rRNA, small cytoplasmic RNA (scRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), tRNA or protein coding regions were discarded. Reads corresponding to annotated 21U-RNAs[8] were also discarded. The remaining aligned reads were then used as guidelines for excising potential miRNA precursor sequences from the genome. Each of these potential precursor sequences were input to the miRDeep algorithm as described above. Scoring of sequences that passed the initial filtering (**Fig. 3**) revealed that 116 sequences passed the cut-off of 1 (all blue, **Fig. 3a**). Of these, 103 were known miRNA precursors (dark blue), corresponding to 102 unique known mature sequences, whereas 13 represented new candidate miRNA precursors, previously unannotated in this species (light blue). Of the 135 known *C. elegans* mature miRNA sequences at miRBase, 115 were present in the data set (**Fig. 4**). Of these, 102 (89%) were successfully recovered by miRDeep (**Fig. 4a**). The total estimated number of false positives was 8 ± 3 (s.d.), corresponding to a signal-to-noise ratio of 15:1 (**Fig. 4b**). The estimated number of false positives for the new predictions was 6.5 ± 2 (s.d.), corresponding to a signal-to-noise ratio of 2:1 (**Fig. 4c**).

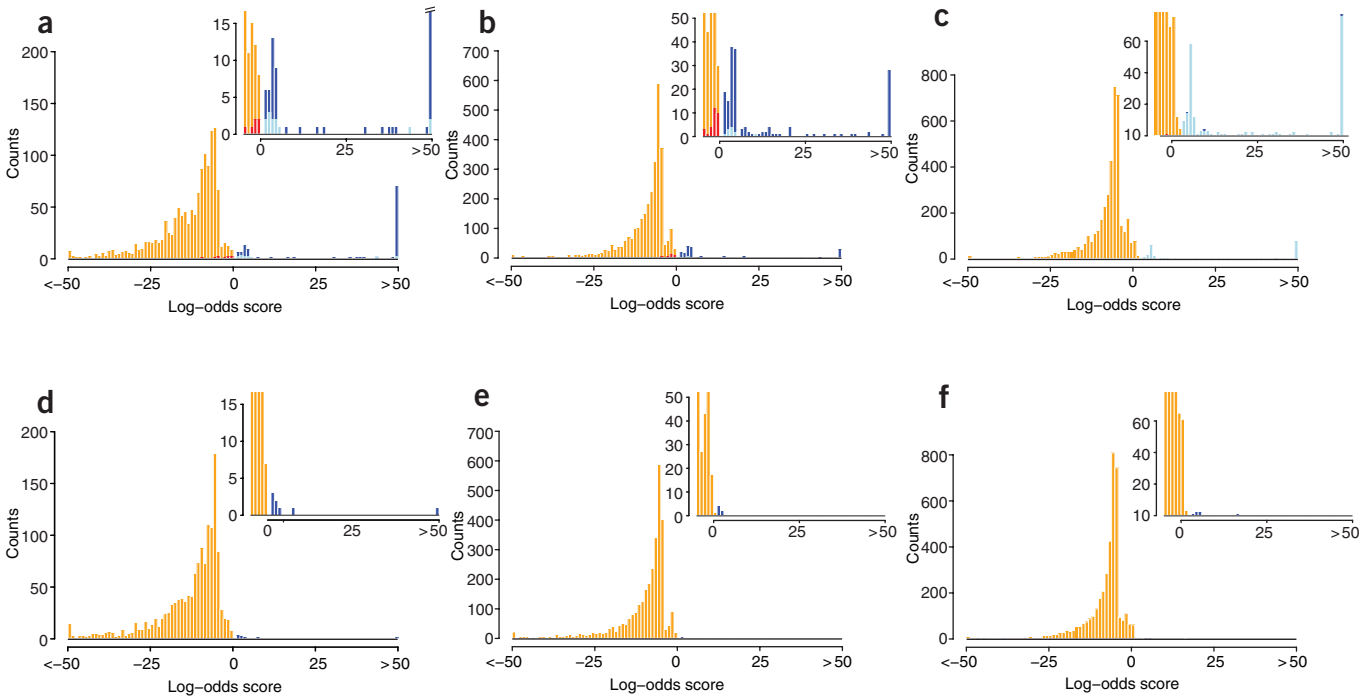


Figure 3 Discovery of known and novel miRNAs by miRDeep. (**a**–**c**) Histograms of miRDeep scores are shown for *C. elegans* (**a**), human (**b**) and dog (**c**) data. The inserts are close-ups. Known miRNA precursors are colored dark blue. False negatives (known miRNA precursors that do not exceed the cut-off of 1 for *C. elegans* and human, or 3 for dog) are plotted in red. Data above the score cut-off are likely novel miRNAs and colored light blue. All other data points are plotted in orange. (**d**–**f**) The statistical controls for *C. elegans* (**d**), human (**e**) and dog (**f**) are shown. Scores exceeding the cut-off are colored in blue (false positives), everything else in orange. These controls show that miRDeep correctly classifies the vast majority of potential miRNA precursors into true miRNAs and likely non-miRNAs, according to our simple model of miRNA biogenesis. The appearance of some false positives with very high scores results from the conservative nature of the statistical controls ('hitchhiking' effect).

Only two more predictions resulted from doing predictions without first discarding reads aligning to known annotations (including 21U-RNAs). This shows that the annotation is not crucial for the prediction accuracy.

The mature and precursor sequences of the 13 novel candidates can be found in the **Supplementary Sequences** online. Eight of the novel miRNAs had 3′ overhangs characteristic of Dicer processing on both hairpin arms (**Supplementary Fig. 1** online). Further, some of the novel miRNA genes had conservation patterns typical for miRNAs (**Supplementary Fig. 2a**,**b** online). Northern blotting confirmed four of the five candidates tested (**Fig. 5**).

These results show, first, that miRDeep can successfully recover known miRNAs with high (89%) sensitivity, second, that miRDeep can successfully discriminate between miRNAs and other types of small RNAs, and finally, that although the data sets used have already been specifically mined for small RNA species[8,12], miRDeep still predicts ten likely novel miRNA genes, while recovering 13 out of 18 precursor candidates predicted previously[8].

A single miRDeep run recovers 28% of known human miRNAs

To produce the second data set, we used the Solexa technology to sequence the small RNA fraction of a human HeLa cell sample. The human genome (~3 Gb) is larger than that of *C. elegans* and has also already been mined extensively for miRNA sequences by conventional cloning of small transcripts, as well as by computational searches and deep sequencing (see, for instance, refs. 5,9,31).

miRNA predictions were made as for the *C. elegans* data set, and reads aligning to annotated rRNA, scRNA, snRNA, snoRNA and tRNA were discarded. In total, 173 sequences passed the cut-off of 1 (all blue, **Fig. 3b**). Of these, 163 were known precursors (dark blue; corresponding to 154 unique known mature miRNA sequences), whereas 10 represented new candidate miRNA precursors (light blue). Sequences of novel candidates are provided in the **Supplementary Sequences**. Further, some of the novel miRNA genes had conservation patterns typical for miRNAs (**Supplementary Fig. 2c**,**d**). Of the 555 known human mature miRNA sequences, 213 were present in the data set. Of these, 154 (72%) were successfully recovered by miRDeep (**Fig. 4d**). The total estimated number of false positives was 6 ± 2 (s.d.), corresponding to a signal-to-noise ratio of 29:1 (**Fig. 4e**). The estimated number of false-positive rates for the new predictions were 5 ± 2 (s.d.), corresponding to a signal-to-noise ratio of 2:1 (**Fig. 4f**).

Thus, despite years of research effort to clone small RNAs in dozens of human tissues, miRDeep recovers 156 (28%) of all known human mature miRNA sequences when analyzing deep-sequencing reads from a single HeLa sample. Perhaps surprisingly, we also found that 213 (~40%) of all known human mature miRNAs can be detected in in our HeLa sample, although roughly half of these are represented by <10 reads.

To summarize, after ~10^6 nonredundant loci were input to miRDeep, the algorithm recovered the majority of the known miRNAs present in the sample, reported ten novel miRNAs and produced only six false positives.

miRDeep discovers >200 dog miRNAs

The third data set was produced by Solexa sequencing the small RNA fraction of a domestic dog lymphocyte sample. Domestic dogs are emerging as an important model system for human disease[32], and are

appealing for miRNA profiling as only six dog miRNA genes are annotated in miRBase[24]. miRNA predictions were made as before, except no reads were discarded based on the annotation. In total, 206 passed the cut-off of 3 (**Fig. 3c**). Of these, 203 represented previously unknown dog candidate miRNA genes (light blue), whereas three represented previously known dog miRNAs (dark blue). As only four known miRNAs are present in the data set, the sensivity is 75% (**Fig. 4g**). The estimated number of false positives both for the total and for the new predictions was 6 ± 2 (s.d.) corresponding to a signal-to-noise ratio of 30:1 (**Fig. 4h,i**). Of the novel miRNAs, 90% had a conserved nucleus sequence (**Supplementary Table 1** online and **Supplementary Sequences**) and 58% had the 3′ overhangs characteristic of Dicer processing. When the novel precursors were compared with known rRNA, scRNA, snRNA, snoRNA, tRNA consensus sequences, only two had any similarity.

Thus, miRDeep can reveal numerous miRNA genes when analyzing data from genomes previously unmined for small RNAs.

Availability of the miRDeep software package

The miRDeep package can be downloaded at http://www.mdc-berlin.de/rajewsky/miRDeep and consists of several specialized Perl scripts that in combination perform the computations described in this study. Beside Perl (available at http://www.perl.com/), the Vienna package[33] (available at http://www.tbi.univie.ac.at/RNA) and the Randfold application[21] (http://bioinformatics.psb.ugent.be/software/details/Randfold) are required dependencies. Also needed is a nucleotide sequence alignment tool such as the NCBI BLAST package[34] (http://www.ncbi.nlm.nih.gov/Ftp/). All of these packages are portable and freely available. As the miRDeep core parameters work independent of species and data sets, no complicated estimation processes are needed. The cut-off can be varied with a single command line argument for custom trade-offs between sensitivity and specificity. The user can choose which potential precursor sequences to input to the core algorithm. These can be either sequences excised from the genome by miRDeep using the aligned reads as guidelines, or custom sequences. After aligning reads to the genome, only a few hours on a standard Linux box are needed for genome-wide prediction using miRDeep.

DISCUSSION

By using a simple model for miRNA precursor processing by Dicer, miRDeep is capable of both recovering the majority of known miRNAs present in heterogeneous deep-sequencing samples and reporting novel miRNAs with high confidence. Estimating the reliability of results by predicting false-positive rates before follow-up experiments is important for most practical applications. Such statistical tests always depend on certain assumptions, but our approach has the virtue of relying on the biological model of miRNA precursor processing by Dicer, which is precisely at the heart of the miRDeep algorithm. Another general limitation of algorithms for miRNA discovery is their reliance on parameters learned from known miRNAs, which introduces bias towards accurate recovery of known miRNAs, but less reliability or sensitivity in discovering novel miRNAs ('overtraining'). However, whereas miRDeep parameters were derived from only a subset of miRNAs, they produce the overall same quality of results when run on very different data sets. Thus, we believe that miRDeep is not overtrained and that it is a widely applicable and flexible tool for researchers wanting to identify known and novel miRNAs in metazoan deep-sequencing samples.

However, to test an extreme case, we ran miRDeep on deep-sequencing data from a planarian sample (unpublished data). Planaria are metazoans, but have roughly equal phylogenetic distance to human and *C. elegans* and reside altogether in a comparatively unexplored branch of the metazoan phylogenetic tree. Sixty-one mature miRNAs had been cloned and sequenced in planaria previously[35]. miRDeep rediscovered 86% of these, while reporting 39 novel miRNAs. Importantly, no genomic annotation information was used. We have validated 16 of 19 tested miRNAs by northern blot analysis (unpublished data). At least 7 out of these 16 miRNAs have not been reported in any other animal, adding confidence to miRDeep results, even in situations where only a minimum of conservation or annotation information is available.

Ruby *et al.*[8] also predicted miRNAs from deep-sequencing data in *C. elegans*, but did not estimate the sensitivity and the false-positive rate of the prediction approach. Although the approach is neither

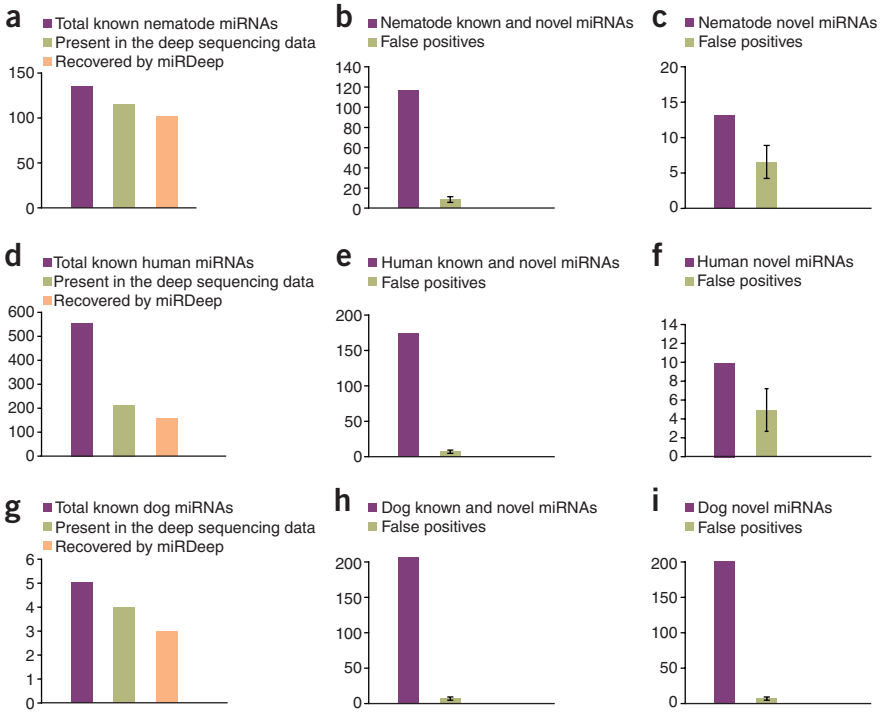


Figure 4 Accuracy of the miRDeep algorithm. (**a**–**i**) The three rows show sensitivity and false-positive estimates for miRDeep results from *C. elegans* (**a**–**c**), human (**d**–**f**) and dog (**g**–**i**). In **a**, **d** and **g**, the total number of mature miRNA sequences known in each species is shown in purple, the total number of mature sequences present in each deep-sequencing data set that matched any of the known mature sequences (allowing for mismatches in the 3′ end) is shown in green and the number of mature sequences recovered in the final set of miRDeep predictions is shown in orange. By this measure, the sensitivity of miRDeep ranges from 72–89%. The false-positive estimations are shown in each data set separately for the total number of miRNA precursor predictions (**b**,**e**,**h**) and for the novel miRNA predictions only (**c**,**f**,**i**). miRNA precursors reported by miRDeep are shown in purple. The estimated number of false positives is shown in green, with error bars indicating the s.d. The signal-to-noise ratios (ratio of the heights of purple and green bars) for total miRNAs range from 15:1 to 30:1. For novel miRNAs, the dog data set has the best quality (signal-to-noise ratio 30:1), as this genome has previously not been mined heavily for miRNAs.

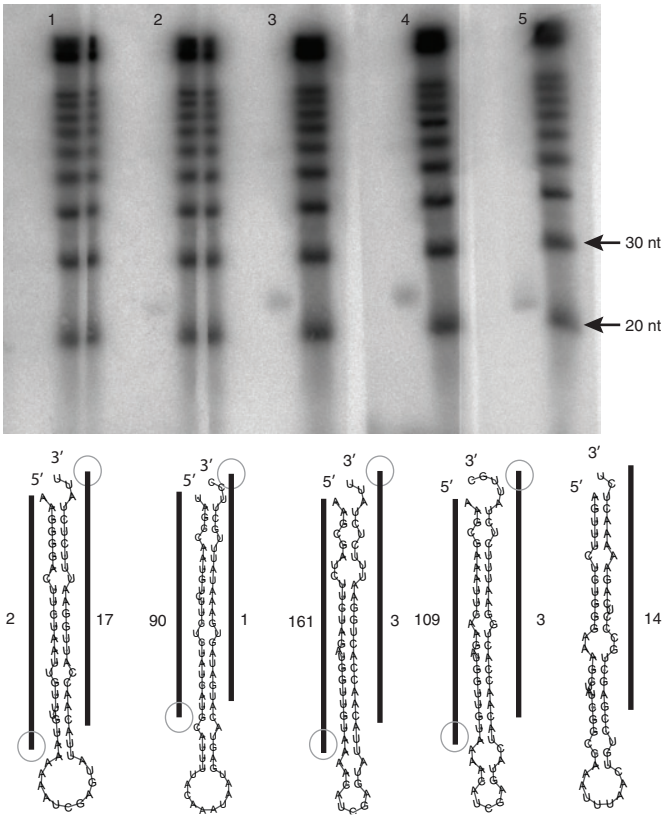


Figure 5 Validating miRDeep candidates by northern blot analysis. Northern blot analysis of five of the novel *C. elegans* miRDeep miRNAs revealed bands corresponding to the mature miRNA product in four out of five candidates (lanes 2–5). The nucleotide length of the mature products as indicated by the RNA marker lanes are consistent with the predicted mature miRNA length in all four cases. The predicted secondary structure of each precursor is provided below. Black vertical bars represent the consensus positions of sequencing reads that mapped to the predicted precursors and numbers indicate the total number of these reads. The gray circles indicate small 3′ overhangs which are known to be typical for Dicer processing.

available as a software package nor described in enough detail to allow us to test their approach on other data sets, running miRDeep on the exact same deep-sequencing data used by Ruby *et al.* recovered 13 out of the 18 novel miRNA precursors predicted by Ruby *et al.*, while reporting 12 additional novel miRNAs. The inclusion of the *C. elegans* deep-sequencing data from Pak *et al.*[12] yielded another novel miRDeep-predicted miRNA.

Berezikov *et al.*[9] described an algorithm that predicts hundreds of novel human miRNA candidates from deep-sequencing data. However, it is difficult to determine how many of these are genuine, as they are typically expressed at extremely low levels. The majority of these candidates are represented only by a single read, making it difficult to decide whether they are genuine miRNAs or degradation products from non-miRNA transcripts. Whereas Berezikov *et al.* estimate that their algorithm predicts one false-positive miRNA for an input of ~100 nonredundant read sequences, miRDeep has several orders-of-magnitude fewer false positives (one false-positive miRNA for every ~25,000 nonredundant read sequences). However, the two algorithms were designed with very different objectives. Whereas the algorithm of Berezikov *et al.* takes the deep-sequencing technology to the limit in terms of sensitivity and seeks to report an exhaustive list of miRNA candidates, miRDeep is designed to recover a large set of real miRNAs in a deep-sequencing sample, while minimizing the number of false positives. It will be interesting to run miRDeep on the data used by Berezikov *et al.* once they are publicly available.

In this study, the potential miRNA precursors that were input to miRDeep were excised from the genomes using the deep-sequencing reads as guidelines, and further filtered by very basic characteristics. We alternatively tried to use candidate miRNA precursor sequences predicted by several advanced miRNA detection algorithms that predict miRNA genes by using support vector machine or other types of learning algorithms based on much more detailed features of miRNA precursor structures (data not shown). However, we found that in all cases this severely compromised the sensitivity of miRDeep without lowering the false-positive rate. Of course, a number of existing miRNA gene-prediction programs have proven to be useful[6]. Therefore, our results suggest that many of these algorithms could potentially be significantly improved by incorporating deep-sequencing data.

METHODS

Preparation of total RNA. Peripheral blood samples were drawn from a male dog (race "Griechischer Laufhund", eleven years old) using a heparin-coated syringe. Following a selective hypotonic lysis of erythrocytes[36], residual white blood cells were collected by centrifugation (500*g*, 5 min, 20 °C), suspended in PBS and immediately used for RNA isolation. Dog total RNA was prepared using the mirVana Isolation Kit (Ambion) according to the manufacturer's protocol. The quality and quantity of resulting total RNA samples was checked using the NanoDrop Spectrometer (ND-1000 Spectrophotometer, Peqlab) as well as the Agilent 2100 Bioanalyzer (RNA Nano Chip, Agilent).

Total RNA was isolated from mixed-stage *C.elegans* population (N2 strain) using TRIZOL reagent (Invitrogen) following the manufacturer's protocol[37]. Total RNA from HeLa cells was also isolated using the TRIZOL protocol.

Northern blots. Validation of miRDeep candidates was done by northern blot analysis as described earlier[38]. Briefly, 90 µg total RNA per lane and a RNA ladder (Decade marker, Ambion) were resolved side by side on a 15% denaturing polyacrylamide gel and transferred onto Hybond-N+ membrane (Amersham, GE Life Sciences). Hybridization and wash steps were performed at 43 °C. The 5′ ^{32}P-radiolabeled oligodeoxynucleotide probes were:

5′-AATAGAGAAATTCCAATGGTTG-3′ for miRDeep-cel-2,
5′-CATGATAGAGAAGACATTGGCTA-3′ for miRDeep-cel-3,
5′-TACAACCATCTAGAAGATCGCTT-3′ for miRDeep-cel-4,
5′-TACAACCATCTTGAATTTCGCTT-3′ for miRDeep-cel-5 and
5′-AGAGTTTTTCTGAGGGCAGCTC-3′ for miRDeep-cel-8.

Solexa sequencing of human and dog small RNAs. Small RNAs from the human and dog total RNA samples were prepared for Solexa sequencing as follows: ~10 µg total RNA were size-fractionated by Novex 15% TBE-Urea gel (Invitrogen) and RNA fragments of length between 20 and 30 bases were isolated. The purified small RNAs were then ligated with 5′ adapter (Illumina). To remove unligated adapters, the ligation products (40–60 bases in length) were gel purified on Novex 15% TBE-Urea gel. Subsequently, the RNA fragments with the adapter at the 5′ end were ligated with 3′ adapters (Illumina). After gel purification on Novex 10% TBE-Urea gel (Invitrogen), RNA fragments with the adapters at both ends (70–90 bases in length) were reverse transcribed and the resulting cDNA was subjected to 15 PCR cycles. The amplification products were loaded on Novex 6% TBE gel (Invitrogen) and the gel band containing 90- to 100-bp fragments was excised. The purified DNA fragments were used directly for cluster generation and 27 (human) or 36 (dog) cycles of sequencing analysis using the Illumina Cluster Station and 1G Genome Analyzer following manufacturer's protocols. Sequencing reads were extracted from the image files generated by Illumina 1G Genome Analyzer using the open source Firecrest and Bustard applications (Illumina).

Obtaining *C. elegans* small RNA 454 sequencing reads. Two published *C. elegans* 454 deep-sequencing data sets were obtained from the GEO database at NCBI.

The first had been produced by sequencing a sample of mixed-stage *C. elegans* fed bacteria that produced double-stranded RNA (accession no. GSE6282). The other had been produced by combining five sequencing reactions of five different mixed-stage samples (accession no. GSE5990).

Aligning the deep-sequencing reads. The deep-sequencing reads of the two *C. elegans* 454 deep-sequencing sets were combined and aligned to the genome (*C. elegans* version ce2, obtained from the UCSC genome database http://genome.ucsc.edu/) using NCBI megablast (BLAST version 2.2.14) with the following options: -W 12 -p 100. Only perfect alignments were retained (full length, 100% identity).

The HeLa cell Solexa data set was aligned to the human genome (*Homo sapiens* version hg18, from UCSC) using megablast, as above. As this data set included adapter sequences, these were subsequently removed using the following approach: alignments were kept that had perfect alignment from nucleotides 1–18, and these alignments were extended until the first mismatch. Any unaligned ends of these reads were assumed to be adapters and were discarded. For each read, alignments of suboptimal length were discarded (if the best alignment was 22 nt, all shorter alignments were discarded).

Adapters were removed from the dog lymphocyte Solexa data set by use of a custom suffix-based mapping tool. First, the adapter sequences were identified in the deep-sequencing reads. We required the presence of minimum 10 nucleotides (nt) of the 5′ adapter sequence with a maximum of three edits (mismatches and/or insertions/deletions). Reads that contained an identified adapter sequence had the adapter removed and were retained, the rest were discarded. The retained reads were mapped to the dog genome (*Canis familiaris* version canFam2, from UCSC) using the custom mapping tool, allowing for up to two edits. For each read, mappings of suboptimal edit distance were discarded (if the best mapping was edit distance 1, all edit distance 2 mappings were discarded).

Excising potential miRNA precursors from the genome using deep-sequencing reads as guidelines. Before excising the potential precursors from the genome using the aligned reads as guidelines, the miRDeep package discards a number of reads unlikely to represent mature miRNA sequences. These reads are only disregarded for purposes of the potential precursor excision, since the total set of reads is used to score the potential precursors (see the next section). More precisely, we discarded reads that aligned to more than five positions in the genome. The vast amount of known mature miRNA reads align to five positions or less (unpublished results), and by discarding reads that align ubiquitously, vast numbers of alignments can be disregarded. Further, *C. elegans* and human reads that overlapped with positions (on either strand) annotated by the UCSC database[39] as rRNA, scRNA, snRNA, snoRNA or tRNA were discarded, as were reads that had perfect alignments to these types of noncoding RNA in the Rfam database[40]. Since it is known that *C. elegans* encodes endogenous small interfering RNAs and 21U-RNAs, all reads overlapping with annotated positions of protein coding sequence or 21U-RNAs[8] were discarded.

The remaining aligned reads were used as guidelines to excise potential precursor sequences from the genome. In the cases where reads aligned to the same strand within 30 nucleotides of each other, they were assumed to represent Dicer products of the same putative miRNA precursor, and were clustered. In these cases, a single sequence, consisting of the clustered region and 25-nucleotide flanks were excised. If such a potential precursor was longer than 140 nucleotides, it was discarded. In the cases where reads aligned more than 30 nucleotides from any other aligned reads on the same strand, two potential precursor sequences of length 110 nt were excised, corresponding to the reads being processed from the right or left arm of a potential precursor sequence.

Probabilistic scoring of the potential miRNA precursors. At this point, potential precursors that did not fold into a hairpin, or that had reads aligning to it in a way that was inconsistent with Dicer processing, were discarded. This was done by a combinatorial investigation of structure and signature. The details are as follows. First, the position of the potential mature miRNA sequence was defined as the position of the most abundant read sequence aligning to the potential precursor sequence. Second, the potential star sequence was defined as the sequence base pairing to the potential mature sequence, correcting for the 2-nt 3′ overhangs. Third, the loop was defined as the sequence between the potential mature and star sequence. Fourth, the potential mature-loop-star structure should form an unbifurcated hairpin, with a minimum of 14 base pairings between the mature and the star sequence. Fifth, for each read it was tested whether it aligned to the potential precursor in consistence with the signature expected from Dicer processing. More precisely, a read is in consistence if it aligns with the potential mature, loop or star, allowing the read to stretch two nucleotides beyond the expected position in the 5′ end or up to five nucleotides in the 3′ end. In the cases where >10% of the reads aligning to a potential precursor were inconsistent with this signature, the potential precursor was discarded. These liberal consistency rules were used to add robustness to the detection of fuzzy endonuclease processing.

Each potential precursor sequence that passed the initial filtering was then scored probabilistically. Our score is the log-odds probability of a sequence being a genuine miRNA precursor versus the probability that it is a background hairpin, given the evidence from the data:

1. score = log (P(pre | data) / P(bgr | data)

The probability of the sequence being a precursor is given by Bayes' theorem:

2. P(pre | data) = P(data | pre) P(pre) / P(data)
3. P(pre | data) = P(abs | pre) P(rel | pre) P(sig | pre) P(star | pre) P(nuc | pre) P(pre) / P(data)

The same holds for the probability of the sequence being a background hairpin:

4. P(bgr | data) = P(data | bgr) P(bgr) / P(data)
5. P(bgr | data) = P(abs | bgr) P(rel | bgr) P(sig | bgr) P(star | bgr) P(nuc | bgr) P(bgr) / P(data)

P(pre) is the prior probability that a potential precursor is actually a miRNA precursor.

P(bgr) is the prior probability that a potential precursor is non-miRNA background hairpin and equal to 1-**P(pre)**.

abs is the estimated minimum free energy of the potential precursor.

P(abs|pre) is the probability that a real miRNA precursor would have the value **abs**.

P(abs|bgr) is the probability that a non-miRNA background hairpin would have the value **abs**.

rel is equal to 1 if the potential precursor sequence is energetically stable, 0 otherwise.

P(rel|pre) is the probability that a real miRNA precursor has the value **rel**.

P(rel|bgr) is the probability that a background precursor has the value **rel**.

sig is the number of reads in the deep-sequencing sample that align to the potential precursor sequence in consistence with Dicer processing (see above).

P(sig|pre) is the probability that a real miRNA precursor has the value **sig** in the deep-sequencing sample.

P(sig|bgr) is the probability that a background hairpin has the value **sig** in the deep-sequencing sample.

star is equal to 0 if the potential precursor sequence has no reads that represent a putative star sequence, and 1 otherwise.

P(star|pre) is the probability that a real miRNA precursor has the value **star** in the deep-sequencing sample.

P(star|bgr) is the probability that a background hairpin has the value of **star** in the deep-sequencing sample.

nuc is an (optional) binary variable. It is 0 if the nt 2–8 from the 5′ end of the putative mature miRNA are not conserved in any other metazoan, and 1 otherwise.

P(nuc|pre) is the probability that a real miRNA precursor has the value of **nuc**.

P(nuc|bgr) is the probability that a background hairpin has the value of **nuc**.

In the above, we are assuming independence between **abs**, **rel**, **sig**, **star** and **nuc**.

Parameter estimation. All parameters were first estimated using *C. elegans* data only:

pre and **brg** are by default set to $P = 0.5$, but can be changed based on the expected miRNA contents in the deep-sequencing samples.

sig. To generate a set of background hairpins, we took the sequences excised from the *C. elegans* genome and discarded the ones that corresponded to known miRNA precursors or that did not have a hairpin structure. The number of remaining hairpins was ~2,000. For each background hairpin, we found the number of reads that aligned perfectly to it. The distribution of these numbers was approximately geometric. The parameter of the geometric distribution (used to model **sig**) was estimated using the mean of the numbers. The same

procedure was used for known *C. elegans* miRNAs to estimate a geometric distribution for real miRNA precursors.

abs. For each background hairpin, the absolute value of the minimum free energy was predicted using RNAfold. The distribution of these values was found to approximate the Gumbel distribution. The parameters for the Gumbel distribution (used to model **abs**) were as estimated in ref. 41. As the Gumbel distribution is a continuous distribution, probabilities were calculated within windows of 1 kcal/mol. The same procedure was used for known *C. elegans* miRNAs to estimate the Gumbel distribution for real miRNA precursors.

rel. A potential precursor was defined to be energetically stable if it had a Randfold $P < 0.05$ (mononucleotide shuffling, 999 permutations). Since it is computationally demanding to produce this large a number of permutations, the contribution of the relative stability to the overall score is only calculated if it can make the difference between the overall score exceeding the cut-off or not. This is the cause of the 'valley' in the score distributions between score 0 and 1 in **Figure 3**.

star is set to 1 if the majority of star reads have a 5′ end that is within one nucleotide of the position expected from Dicer processing (taking into account 3′ overhangs).

For both true precursor hairpins and background hairpins, the probabilities for **rel**, **star** and **nuc** were set according to raw relative frequencies. If, for instance, 1% of the background hairpins had a conserved nucleus, **P(nuc|bgr)** would be set to 0.01.

In some samples, we observed that many known small RNAs other than miRNAs are transcribed in large numbers from a single locus from one strand only. Therefore, we limited the contribution of **sig** to the total score to 0 unless the star sequence is represented by at least one read. In practice this means that the structure scoring of **abs** and **rel** becomes more important when the deep-sequencing data are ambiguous.

The entire parameter estimation procedure was repeated in planaria, using the known precursors of the planarian *Schmidtea mediterranea* (also from miRBase) and unpublished planarian 454 data. Although *C. elegans* and planarians are separated by a large phylogenetic distance, the parameter estimates were similar, suggesting that the estimation process is largely species-independent. The pooled training sets of these two species have been used to estimate the final parameter set for the current study.

Controls. The number of known mature miRNA sequences present in the data sets was estimated by finding how many mature sequences aligned perfectly to the deep-sequencing reads, allowing for mismatches in the last three nucleotides of the mature sequences. This was done on the raw deep-sequencing data sets, just after adapters had been removed. The number of known mature miRNA sequences in the predictions was estimated by finding how many mature sequences aligned perfectly to the final set of predicted miRNA precursors. The sensitivity was estimated as the 'number of mature miRNA sequences recovered' divided by the 'number of mature sequences present in the data set'. Both when making the controls and when making the actual predictions, special care was taken to ensure that no miRNAs were scored higher because the sequence of the miRNA was included in the conservation set (circular inference).

The false-positive rate was estimated using a permutation approach. For each potential precursor sequence, the protocol generates a secondary structure prediction and a processing signature containing information on the positions and frequencies of aligned reads. The controls were made such that all structures and signatures were maintained, but the structure and signature pairings were permuted. In all other respects, the runs were performed as described above. For each estimation of the false-positive rate, 100 independent permutations were used.

Comparing novel dog miRNA precursors to Rfam sequences. The set of novel dog miRNA precursor candidates were aligned against the full set of noncoding sequences obtained at Rfam using NCBI blastn with the following options: -F F –e 1e-5. Only two of the candidates had any similarity to non-miRNA sequences (these were snoRNA sequences).

Contribution of scored features to overall accuracy. To assess the contribution of the scored features to the accuracy, we ran miRDeep on the human data, systematically omitting parts of the algorithm. In some cases it is not transparent if changes in sensitivity and false-positive rate actually improve or worsen the algorithm (for instance, when both sensitivity and false-positive rate go up). Therefore the score cut-off was varied in each run such that the sensitivity remained constant (at 72%). We then recorded the change in false positives. Each run was repeated ten times and the mean number of false positives noted. For example, we found that omitting the hairpin stability scoring with Randfold boosted the false-positive rate on average by a factor of 1.9. We found in all cases that the elimination of a score feature increased the number of false positives (minimum free energy 2.2, star sequence 3, conservation 3). Omitting all four score features increased the number of false positives by a factor of 17. Additionally allowing nonhairpins boosted the number of false positives by a factor of 42.

This shows that all features scored by miRDeep significantly contribute to the accuracy. Individual score features can in most cases be omitted, since an increase by a factor of two or three in the false-positive rate can often be tolerated. This means, for instance, that the computational speed of miRDeep can be substantially increased through omission of the Randfold scoring. It also means that conservation scoring can be omitted. However, when miRDeep is run on already mined data, or in genomes that have been heavily mined for small RNAs, we recommend that all parts are included to get the highest possible signal-to-noise ratio for the novel predictions.

The miRDeep software package. The miRDeep software package consists of seven documented Perl scripts that should be run sequentially by the user. miRDeep can be run on Linux or Windows platforms or any other system that supports Perl.

1. **blastoutparse.pl** is used to parse standard NCBI BLAST output format into a custom tabular separated format ('blastparsed').
2. **blastparselect.pl** cleans the output from blastoutparse.pl.
3. **filter_alignments.pl** filters the alignments of deep-sequencing reads to a genome. It filters when only a limited part of a read is aligned. It can also filter reads that are aligning multiple times (user-specified) to the genome. The basic input is a file in blastparsed format.
4. **overlap.pl** can be used (user specified) to remove reads that align to the genome in positions that overlap with selected annotation tracks provided by the user (e.g., known rRNAs, tRNAs). The basic input is a file in blastparsed format and an annotation file in standard gff format.
5. **excise_candidate.pl** cuts out potential precursor sequences from a genome using aligned reads as guidelines. The basic input is a file in blastparsed format and a genome FASTA file. The basic output is also FASTA format.
6. **mirdeep.pl** is the core algorithm. Several files are given as input. The first is a file in blastparsed format giving information on reads aligning to the potential precursors. The second is an RNAfold output file giving information on the sequence, structure and absolute stability of the potential precursors. Several command line options are available. One option inputs a FASTA file containing known mature miRNA sequences to allow for conservation scoring. Another option allows for a sensitive run optimized for Sanger sequences obtained through conventional small RNA cloning. Another option evaluates Drosha stem recognition by scoring the number of base pairings formed by the sequences immediately flanking the potential precursor sequence. A further option uses the Randfold algorithm to score the relative stability of potential precursors that have a score close to the set cut-off. Basic output of the algorithm is the total information on the predicted miRNA precursors, including structure prediction, minimum free energy, signature and the scoring contributions of all evaluated features.
7. **permute_structure.pl** permutes the id and sequence/structure combinations of an RNAfold output file. This is used to do the permutation controls.

Accession codes. NCBI Gene Expression Omnibus (GEO). Data sets have been deposited with accession codes GSE10825 and GSE10829.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank H.-H. Ropers for making possible the deep sequencing of HeLa cell and the dog lymphocyte RNA at the Max Planck Institute for Molecular Genetics in Berlin. We are indebted to Alejandro Sánchez Alvarado and John Kim for the planarian data. Thomas Isenbarger helped at the very initial stage of the project. Eugene Berezikov kindly provided unpublished deep-sequencing data (not used in this study). Ralf Bundschuh helped with parameter estimations. M.R.F. acknowledges a fellowship from the Max Delbrück Center. J.M. acknowledges a fellowship from Deutsche Forschungsgemeinschaft (International Research

Training Group 1360). Finally, many thanks to the members of the Rajewsky lab for countless hours of stimulating discussions, and in particular to Nadine Thierfelder and Svetlana Lebedeva for providing the HeLa cell and *C. elegans* samples.

1. Bushati, N. & Cohen, S.M. microRNA Functions. *Annu. Rev. Cell Dev. Biol.* **23**, 175–205 (2007).
2. Bartel, D.P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
3. Miranda, K.C. *et al.* A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. *Cell* **126**, 1203–1217 (2006).
4. Aravin, A. & Tuschl, T. Identification and characterization of small RNAs involved in RNA silencing. *FEBS Lett.* **579**, 5830–5840 (2005).
5. Landgraf, P. *et al.* A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).
6. Bentwich, I. Prediction and validation of microRNAs and their targets. *FEBS Lett.* **579**, 5904–5910 (2005).
7. Lau, N.C. *et al.* Characterization of the piRNA complex from rat testes. *Science* **313**, 363–367 (2006).
8. Ruby, J.G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans. Cell* **127**, 1193–1207 (2006).
9. Berezikov, E. *et al.* Diversity of microRNAs in human and chimpanzee brain. *Nat. Genet.* **38**, 1375–1377 (2006).
10. Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K. & Hannon, G.J. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**, 744–747 (2007).
11. Girard, A., Sachidanandam, R., Hannon, G.J. & Carmell, M.A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199–202 (2006).
12. Pak, J. & Fire, A. Distinct populations of primary and secondary effectors during RNAi in *C. elegans. Science* **315**, 241–244 (2007).
13. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila. Cell* **128**, 1089–1103 (2007).
14. Houwing, S. *et al.* A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**, 69–82 (2007).
15. Tarasov, V. *et al.* Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G1-arrest. *Cell Cycle* **6**, 1586–1593 (2007).
16. Chen, K. & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.* **8**, 93–103 (2007).
17. Grishok, A. *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**, 23–34 (2001).
18. Hutvagner, G. *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834–838 (2001).
19. Filipowicz, W., Bhattacharyya, S.N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.* **9**, 102–114 (2008).
20. Rajewsky, N. microRNA target predictions in animals. *Nat. Genet.* **38** Suppl, S8–S13 (2006).
21. Bonnet, E., Wuyts, J., Rouze, P. & Van de Peer, Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**, 2911–2917 (2004).
22. Pasquinelli, A.E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86–89 (2000).
23. Chen, K. & Rajewsky, N. Deep conservation of microRNA-target relationships and 3′UTR motifs in vertebrates, flies, and nematodes. *Cold Spring Harb. Symp. Quant. Biol.* **71**, 149–156 (2006).
24. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. & Enright, A.J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34**, D140–D144 (2006).
25. Berezikov, E. *et al.* Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.* **16**, 1289–1298 (2006).
26. Vagin, V.V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320–324 (2006).
27. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in *C. elegans. Cell*, **75**, 855–862 (1993).
28. Lee, R.C., Feinbaum, R.L. & Ambros, V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843–854 (1993).
29. Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in *C. elegans. Curr. Biol.* **13**, 807–818 (2003).
30. Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P. & Burge, C.B. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* **10**, 1309–1322 (2004).
31. Berezikov, E. *et al.* Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120**, 21–24 (2005).
32. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
33. Hofacker, I.L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429–3431 (2003).
34. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
35. Palakodeti, D., Smielewska, M. & Graveley, B.R. MicroRNAs from the Planarian Schmidtea mediterranea: a model system for stem cell biology. *RNA* **12**, 1640–1649 (2006).
36. Rettig, M.P. *et al.* Evaluation of biochemical changes during in vivo erythrocyte senescence in the dog. *Blood* **93**, 376–384 (1999).
37. Chomczynski, P. & Sacchi, N. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal. Biochem.* **162**, 156–159 (1987).
38. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853–858 (2001).
39. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).
40. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
41. Altschul, S.F., Bundschuh, R., Olsen, R. & Hwa, T. The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res.* **29**, 351–361 (2001).

High-resolution profiling and discovery of planarian small RNAs

Marc R. Friedländer[a,1], Catherine Adamidi[a,1], Ting Han[b], Svetlana Lebedeva[a], Thomas A. Isenbarger[c], Martin Hirst[d], Marco Marra[d], Chad Nusbaum[e], William L. Lee[e], James C. Jenkin[f], Alejandro Sánchez Alvarado[f], John K. Kim[b], and Nikolaus Rajewsky[a,2]

[a]Max Delbrück Centrum für Molekulare Medizin, Robert-Rössle-Strasse 10, D-13125 Berlin-Buch, Germany; [b]Department of Human Genetics, Life Sciences Institute, University of Michigan, Ann Arbor, MI 48109; [c]Departments of Bacteriology and Plant Pathology, University of Wisconsin, 1550 Linden Drive, Madison, WI 53706-1521; [d]Genome Sciences Centre, British Columbia Cancer Center, 675 West 10th Avenue, Vancouver, BC, Canada V5Z 1L3; [e]Broad Institute of Massachusetts Institute of Technology and Harvard University, 320 Charles Street, Cambridge, MA 02141; and [f]Department of Neurobiology and Anatomy, Howard Hughes Medical Institute, University of Utah School of Medicine, 401 Medical Research Education Building, 20 North 1900 East, Salt Lake City, UT 84132

Communicated by Gary Ruvkun, Massachusetts General Hospital, Boston, MA, May 15, 2009 (received for review March 17, 2009)

Freshwater planarian flatworms possess uncanny regenerative capacities mediated by abundant and collectively totipotent adult stem cells. Key functions of these cells during regeneration and tissue homeostasis have been shown to depend on PIWI, a molecule required for Piwi-interacting RNA (piRNA) expression in planarians. Nevertheless, the full complement of piRNAs and microRNAs (miRNAs) in this organism has yet to be defined. Here we report on the large-scale cloning and sequencing of small RNAs from the planarian *Schmidtea mediterranea*, yielding altogether millions of sequenced, unique small RNAs. We show that piRNAs are in part organized in genomic clusters and that they share characteristic features with mammalian and fly piRNAs. We further identify 61 novel miRNA genes and thus double the number of known planarian miRNAs. Sequencing, as well as quantitative PCR of small RNAs, uncovered 10 miRNAs enriched in planarian stem cells. These miRNAs are downregulated in animals in which stem cells have been abrogated by irradiation, and thus constitute miRNAs likely associated with specific stem-cell functions. Altogether, we present the first comprehensive small RNA analysis in animals belonging to the third animal superphylum, the Lophotrochozoa, and single out a number of miRNAs that may function in regeneration. Several of these miRNAs are deeply conserved in animals.

microRNAs | miRNAs | piRNAs | regeneration | stem cells

Planarians have become a molecularly tractable model system in which to study regeneration, tissue homeostasis, and stem-cell biology (1). Planaria are free-living, triploblastic flatworms of the phylum Platyhelminthes, which is presently considered to belong to the superphylum Lophotrochozoa. Model systems for modern molecular and developmental biology have almost exclusively focused on the other 2 superphyla, i.e., the Deuterostomes (which includes vertebrates) and the Ecdysozoa (e.g., *Caenorhabditis elegans* and *Drosophila melanogaster*). Unlike these model systems, planarians possess remarkable regeneration abilities. Decapitation, for example, results in the complete regeneration of the head within 7 days after amputation. Such robust restoration of missing body parts is mediated by adult stem cells known as neoblasts (2). Of the thousands of known planarian species, *Schmidtea mediterranea* is arguably the species of choice for modern molecular biology and high-throughput, genome-wide approaches because it is diploid, it exists in sexual and asexual strains, and its genome has recently been sequenced and annotated (3). The size of its genome is roughly a third of the human genome, and ≈80% of the ≈20,000 annotated planarian genes have orthologs in humans. Moreover, by morphology alone, neoblasts and their immediate division progeny comprise ≈25% of all cells in the adult animal (4). In addition, RNAi screens have identified hundreds of genes specifically linked to planarian regeneration and stem-cell biology (5). Many of these genes are conserved in humans, and thus understanding planarian regeneration promises to yield important insights into human regeneration and stem cell biology.

In recent years, small, noncoding RNAs have emerged as essential players in almost all biological processes. Many different animal small-RNA species have by now been identified, although the biological functions of these species remain largely unclear (6, 7). Important exceptions are microRNAs (miRNAs) and Piwi-interacting RNAs (piRNAs). miRNAs have been shown to play important roles in many differentiation processes, including regeneration (8), whereas at least one function of piRNAs has been shown to be in maintaining the integrity of the germ line (6). PIWI proteins are essential for the biogenesis and function of piRNAs, and they appear to have undergone an expansion in the planarian genome. We have identified at least 7 likely planarian PIWI genes, of which 3 (SMEDWI-1–3) have been in part functionally characterized (9, 10). For example, depletion of SMEDWI-2 has been shown to generate specific defects in stem-cell-mediated regeneration and homeostasis (9). Because neoblasts can give rise to germ-line cells in planaria, it is perhaps not surprising that at least SMEDWI-1 and SMEDWI-2 proteins are specifically expressed in neoblasts, and that depletion of SMEDWI-2 or SMEDWI-3 reduces piRNA production and both are required for neoblast function and regeneration (10).

Given the importance of miRNAs and piRNAs for planarian and stem-cell biology, it is essential to identify and classify small RNAs in *S. mediterranea*. Presently, 63 planarian miRNA genes encoding for 61 unique, mature miRNAs have been identified (11) and attempts have been made to describe their expression mainly by in situ hybridization of primary miRNA transcripts (12). However, mature miRNA expression can be highly regulated (13). Therefore, to determine the definitive spatial distribution of miRNAs, expression patterns of primary transcripts have to be complemented by mature miRNA expression data. Additionally, all known planarian miRNAs have been identified by classic cloning and Sanger sequencing, and it is highly likely that the true number of planarian miRNAs is much higher. A recent study (10) has further identified a few thousand piRNAs, which is also almost certainly a vast underestimate of the true number of planarian piRNAs (14).

We thus used massive, next-generation sequencing methods to define the full complement of small RNAs present in neoblasts,

Author contributions: M.R.F., C.A., A.S.A., J.K.K., and N.R. designed research; M.R.F., C.A., T.H., S.L., M.H., M.M., C.N., W.L.L., J.K.K., and N.R. performed research; M.R.F., C.A., T.H., S.L., T.A.I., M.H., M.M., C.N., W.L.L., J.C.J., A.S.A., J.K.K., and N.R. contributed new reagents/analytic tools; M.R.F., C.A., and N.R. analyzed data; and M.R.F., C.A., and N.R. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nln.nih.gov/geo (accession no. GSE16159).

[1]M.R.F. and C.A. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: rajewsky@mdc-berlin.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0905222106/DCSupplemental.

Table 1. The 6 deep-sequencing datasets derived from untreated and irradiated planarians and isolated neoblasts

Datasets	Sequencing platform	Sample type	Number of unique mapped reads	Number of mapped reads	Number of loci
1	454	Neoblast	25,256	63,278	95,951
2	454	Untreated	27,461	93,412	104,769
3	454	Irradiated	21,163	61,391	85,051
4	Solexa	Neoblast	86,063	91,371	417,646
5	Solexa	Untreated	984,459	1,784,859	4,064,381
6	Solexa	Irradiated	767,496	2,050,669	3,018,709
All	All	All	1,507,162	4,144,980	6,700,894

animals depleted of neoblasts, and whole animals. Altogether, we cloned, sequenced, mapped, and annotated millions of small RNAs. Extensive computational, qPCR, and Northern analyses allowed us to double the number of known planarian miRNAs, quantify their expression, and identify a number of mature miRNAs likely to be involved in stem-cell biology. Furthermore, we were able to study the expression, genomic organization, and biogenesis features of planarian piRNAs at a resolution orders of magnitude higher than any previous studies. Our dataset allowed us to compare planarian piRNA characteristics with known piRNA features in mammals and ecdysozoans. Altogether, our work brings the characterization and annotation of small RNAs in planarians to a depth that is at par with other model systems such as *C. elegans*.

Results

Comprehensive and Quantitative Deep Sequencing of Planarian Small RNAs. To profile expression differences of small RNAs, we wished to compare neoblasts, intact animals, and animals devoid of neoblasts with each other. Therefore, RNA was obtained from the clonal asexual strain CIW4 of *S. mediterranea* from FACS-purified neoblasts, intact animals, and irradiated animals in which neoblasts were eliminated by radiation (1). Each of the 3 samples was sequenced with 2 different methods. Solexa (Illumina) technology was used to profile all species of small RNAs (size selection: 18–40 nt). Furthermore, we used the 454 Life Sciences (Roche) technology to specifically profile Dicer products (such as miRNAs) by using a more narrow size selection of 18–25 nt. By using a stringent mapping procedure (see *SI Text*), we matched a total of ≈4.2 million sequencing reads to ≈6.7 million loci in the planarian genome. Table 1 gives an overview of the 6 deep-sequencing datasets. We next assessed the samples' quality by 3 criteria: coverage, reproducibility, and accuracy of expression quantitation.

Coverage. To estimate the coverage of planarian small RNAs by the sequenced RNAs, we computed the overlap of our mapped reads with known planarian miRNAs and piRNAs. Previously identified miRNAs were detected by conventional cloning and sequencing small RNAs from *S. mediterranea* whole-body samples with a median miRNA count of 4 (11). We found all of these miRNAs in our pooled datasets, with a median of >9,000 counts (Table S1). Furthermore, our data contain the lowly expressed "star" miRNAs for 62 of the 63 miRNA genes. Another recent study reported ≈4,800 planarian piRNAs deep-sequenced from whole-body samples of planarians (10). We found 38% of these piRNAs in our data. Considering that animal piRNA populations are estimated to consist of hundreds of thousands of unique sequences (14), it is not surprising that our sequencing of piRNAs is not fully saturated.

Reproducibility. We compared 2 Solexa datasets obtained by sequencing biological replicate planarian samples. For each planarian miRNA in miRBase, we plotted the number of times the miRNA was sequenced in one sample vs. the other sample (Fig. 1*A*). The correlation was almost perfect (Pearson's correlation = 0.996), indicating high reproducibility.

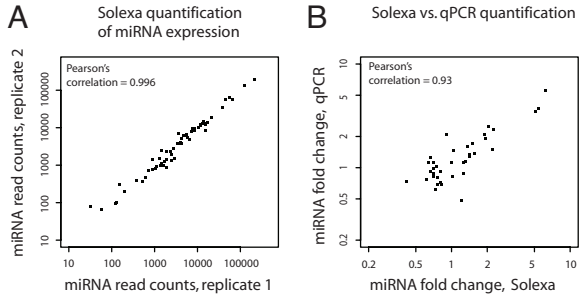


Fig. 1. Reproducible and quantitative sequencing of small RNAs. (*A*) Reproducibility of Solexa quantitation of miRNA expression. Each data point represents 1 miRNA. Samples are independent biological replicates of irradiated planarians. (*B*) Solexa vs. qPCR quantitation of miRNA fold-changes. Each data point represents 1 miRNA. Independent biological replicates were used for the Solexa and the qPCR quantitation.

Accuracy of Expression Quantitation. We investigated whether our deep-sequencing data can accurately quantify differential miRNA expression. We measured expression fold-changes between intact and irradiated samples for 35 planarian miRNAs by using our Solexa data and quantitative PCR in samples from independent biological replicates (Taqman assay; *Methods*). We found a strong correlation between the deep-sequencing data and the qPCR measurements (Fig. 1*B*, Pearson's correlation = 0.93). We conclude that our data are comprehensive, reproducible, and can be used to quantify miRNA expression across samples.

Planarian Small RNAs Are Predominantly miRNAs and piRNAs. We next identified the types of small RNAs present in planarians and quantified their expression in neoblast vs. whole-body extracts. We hypothesized that the comparison of neoblasts with an untreated whole-body sample should identify small RNA species up-regulated in the planarian adult stem cells. If such species are in fact specific to neoblasts, we would further expect them to have reduced expression in the irradiated whole-body sample compared with the samples from the intact, unirradiated animals. Moreover, these comparisons would allow us to detect artifacts, i.e., highly expressed small RNAs, that may have arisen as a result of cell dissociation and/or cell sorting.

Small RNAs in the untreated sample showed a bimodal length distribution with 2 distinct peaks at nucleotides 22 and 32 (Fig. 2*A*). We first selected reads that mapped to known planarian miRNAs from miRBase as well as our novel miRNAs (see below). The length distribution of these reads had a single peak at nucleotide 22, typical for miRNAs (Fig. 2*B*). In fact, these 122 miRNAs (known and novel) account for the entire 22-nt peak in Fig. 2*A*, suggesting that few miRNAs remain to be discovered in *S. mediterranea*. When subtracting all reads mapping to annotated miRNAs, rRNAs and tRNAs, and coding exons, the length distribution forms a distinct peak at nucleotide 32 (Fig. 2*C*). We tentatively refer to these sequences as piRNAs, and will present further evidence for this classification in the next section on piRNAs. Reads mapping to annotated coding exons display a clear peak approximately at nucleotide 32 (Fig. 2*D*). However, this is likely an artifact caused by ambiguous read-mappings and the genome annotation (for discussion see *SI Text* and Table S2).

We next estimated the relative abundance of different classes of small RNAs across the 3 sample types (Fig. 2 *E*–*G*, pie charts). The intergenic piRNA fraction is predominant in sorted neoblasts (82%), intermediate in the untreated sample (61%), and low in the irradiated sample (25%). The increased fractions of rRNA and

BIOCHEMISTRY

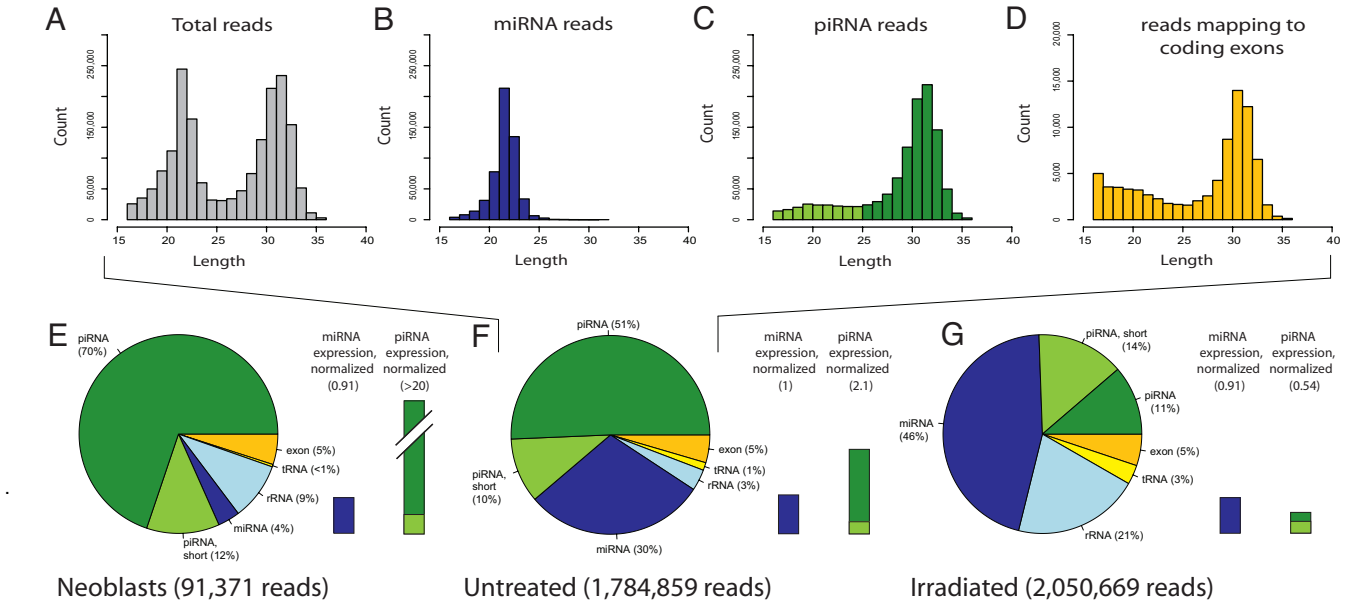


Fig. 2. Small RNA contents in planarian neoblasts and whole-body samples. (*A–D*) Length profiles of Solexa reads from sequencing of the untreated whole-body sample. Only reads successfully mapped to the genome are considered. (*A*) All reads, (*B*) miRNA reads, (*C*) piRNA reads, and (*D*) reads mapping to coding exons are shown. Note that *D* is on a different scale. piRNA reads >25 nt (dark green) have features characteristic for piRNAs (see sections on piRNAs). These features were still present but weaker (70–80%, see *SI Text*) for the remaining piRNA reads (light green, length 17–25 nt). (*E–G*, pie charts) Contents of RNA species in the three Solexa datasets: (*E*) neoblasts, (*F*) untreated planarians, and (*G*) irradiated planarians. Dark green, piRNAs; light green, short piRNAs; dark blue, miRNAs; light blue, rRNA; yellow, tRNA; orange, mRNA. (*E–G*, bar graphs) miRNA (blue), piRNA (dark green), and short piRNA (light green) expression. The 454 data, produced by sequencing RNAs that were 18–25-nt long, were used to estimate miRNA expression, whereas the Solexa data were used for piRNA expression. Total miRNA and piRNA read counts were normalized to miR-71c. miRNA expression of the untreated sample was set to 1, and the other expression bars were scaled accordingly (numbers in parentheses above each bar). Note that piRNA expression in the neoblast sample is out of scale.

short piRNAs in the irradiated sample could be a result of degradation. In contrast, the miRNA fraction is low in the neoblast sample (4%), intermediate in the untreated sample (30%), and larger in the irradiated sample (46%).

Comparing the abundance of each class of small RNAs across different samples requires normalizing contents to a stably expressed endogenous control. We used miR-71c for library normalization because we observed that this miRNA is robustly and constantly expressed across our 3 samples based on a quantitative Taqman assay (*SI Text* and Fig. 3). By normalizing the total read counts of miRNAs and piRNAs to the read count of miR-71c, we were able to estimate the relative expression of small RNAs across samples (Fig. 2 *E–G*, bar graphs). Intergenic piRNAs have very high expression in neoblasts (>10-fold higher than in untreated whole-body planarians) and low expression in irradiated planarians (4-fold lower than in untreated planarians), consistent with the idea that piRNAs may be up-regulated in neoblasts and their division progeny, i.e., where PIWI proteins are specifically expressed (10). In contrast, total miRNA contents appeared roughly constant over the 3 samples, although the abundances of individual miRNAs varied. We independently repeated this analysis with 2 other miRNAs (miR-36 and miR-36c) that appeared roughly constant and obtained comparable results (see Table S3).

Planarian piRNAs Share Key Features with Mammalian and Fly piRNAs. To characterize planarian piRNAs, we analyzed deep-sequencing reads that did not map to annotated miRNAs, rRNAs, tRNAs, or coding sequences (Fig. 2*C*). These reads display 2 of the defining features of mammalian and fly piRNAs (reviewed in ref. 14): a length distribution peaking approximately at nucleotide 30 and a diverse population (≈1.2 million unique sequences in our Solexa data). Northern blots validated size and expression of 3 planarian piRNAs (Fig. 4*A*).

Planarian piRNAs Display a Clear Tendency to Overlap by 10 Nucleotides. The current model of biogenesis proposes that piRNAs are generated through iterative PIWI-mediated cleavage of transcripts with complementary sequence [the "ping-pong" amplification mechanism (15, 16)]. According to this model, piRNAs that map to opposite genomic strands tend to overlap by 10 nt. We investigated whether this signature is conserved in planarians. However, this is difficult because many piRNAs map to numerous genomic loci. For instance, if 2 reads map to the same 100 loci, their overlap would

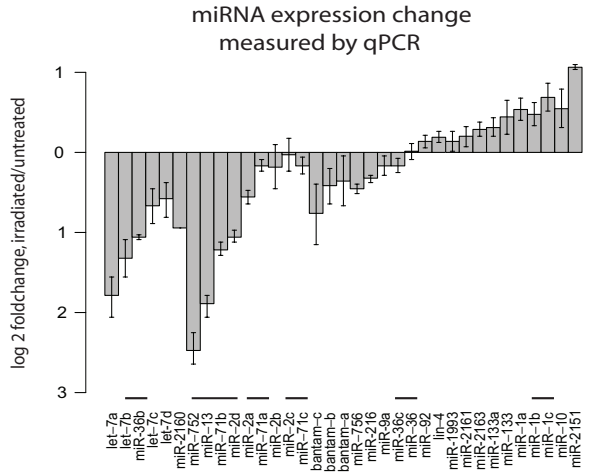


Fig. 3. miRNA expression fold-changes measured by qPCR. Total RNA from untreated and irradiated animals was used to quantify expression fold-changes of 35 miRNAs by qPCR. Data are relative to expression detected for the ubiquitously expressed control *ura4* (see *Methods*). Each miRNA is grouped with its family members. Horizontal bars indicate miRNA gene clusters.

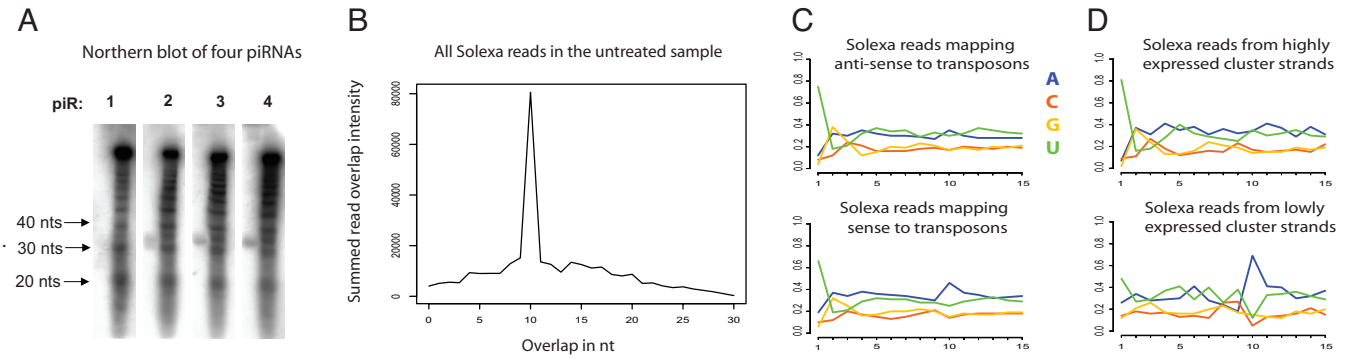


Fig. 4. Features of planarian piRNAs. (*A*) Northern blot analysis of 4 annotated piRNAs. Bands ≈32-nt long are visible for 3 of these piRNAs. (*B*) Summed overlap intensities for Solexa reads in the untreated sample. The horizontal axis is the length of overlap in nucleotides between the 5′ ends of reads mapping to the same genomic locus on opposite strands. The vertical axis shows the intensity of the overlap, summed over the entire dataset. (*C* and *D*) Sequence biases of piRNAs. The horizontal axis shows nucleotide positions from the 5′ end, the vertical axis represents nucleotide fractions. (*E*) piRNAs mapping antisense (*Top*) and sense (*Bottom*) to transposons. (*F*) piRNAs from highly (*Top*) and lowly (*Bottom*) expressed cluster strands.

be counted 100 times. Thus, to avoid potentially inflated counts, we assigned "intensities" to mappings that were inverse to the number of mappings for the read. For example, a read mapping to 10 loci would be assigned an intensity of 0.1. Summing overlap intensities over each of our datasets, yielded major peaks with an overlap of exactly 10 nt in the neoblast and the untreated sample (Fig. 4*B*), whereas the peak for the irradiated sample was greatly reduced (Fig. S1).

Planarian Primary piRNAs Map Antisense to Transposons. An important function of mammalian and fly piRNAs is to silence transposons. In mouse testes, PIWI proteins cleave transposon mRNA to generate primary piRNAs, with a uracil in the 5′ end (17). Primary piRNAs base pair with long transcripts that contain complementary sequence to cleave out secondary piRNAs, which thus typically have an adenosine at position 10. In fly testes, this is reversed. Primary piRNAs are cleaved from transcripts antisense to transposons, and the secondary piRNAs are cleaved from the transposon mRNA (15, 16).

We found that 32% of the planarian piRNAs map to annotated transposons (*SI Text*). piRNAs mapping antisense to transposons have a clear tendency for a beginning uracil and no other sequence biases (Fig. 4*C*), indicating that these are primary piRNAs. piRNAs mapping in the sense orientation to transposons have a bias toward a beginning uracil and an adenosine at position 10.

Planarian piRNAs Locate to Transposons as much as Mouse Prepachytene piRNAs. Mammalian and fly piRNAs differ on the fraction of the population that is transcribed from transposons. Mouse pachytene piRNAs have no reported role in transposon silencing, and map to mouse transposons less than would be expected from the transposon genome coverage. In contrast, mouse prepachytene piRNAs and fly piRNAs have reported roles in transposon silencing (reviewed in ref. 18). These map to transposons as much (mouse prepachytene) or more (fly) than would be expected from the transposon genomic coverage.

Planarian transposons cover 31% of the genome and 32% of the piRNAs. These numbers resemble mouse prepachytene piRNAs. However, planarian piRNAs do display biases toward particular classes of transposons. For instance, Mariner elements, active in planarians (19), have 1.8 times more piRNAs mapping than would be expected by chance, whereas PiggyBac have half the number of mapping piRNAs as would be expected (Table S4). These findings are significant ($P \approx 0$; see *SI Text*). piRNA transposon association changes little across the 3 samples.

Planarian piRNA Clusters Display Strand Expression Bias but Seem Not to Resemble Master Loci. We observed that planarian piRNAs, similar to those of mammals and flies, tend to map to discrete regions. To annotate these piRNA clusters, we located 10-kb regions of the genome to which 100 or more long piRNAs can be unambiguously traced and where instances of 10-nt overlaps between such piRNAs occur. This yielded 119 piRNA cluster candidates to which 6% of all planarian piRNAs in the untreated sample can be traced (Table S5). These clusters are thus highly (and significantly; see *SI Text*) enriched in piRNAs, given that they only constitute about one thousandth of the planarian genome.

The majority (92%) of planarian piRNA clusters displayed a strong strand bias, with piRNA mapping intensities 10 times or higher on one strand (see Fig. S2). piRNAs originating from highly expressed cluster strands, like primary piRNAs in mouse and fly, have a strong bias for a 5′ uracil, whereas the ones from the lowly expressed cluster strands, like secondary piRNAs, have a strong tendency for an adenosine at position 10 (see Fig. 4*D*). Similar strand expression biases are observed in the fly "master loci", which are piRNA clusters densely packed with nonfunctional transposons. However, we did not identify any master loci in the planarian genome, as none of the piRNA clusters contained large numbers of transposons.

Discovery and Validation of Novel Planarian miRNAs. To discover novel miRNAs, we used miRDeep, an algorithm to detect and score Dicer hairpin products such as miRNAs in deep-sequencing data (20). Varying score cut-offs allow trade-offs between sensitivity and specificity. Sensitivity is computed as the fraction of known miRNAs recovered, whereas false positives are estimated by stringent statistical controls (20).

We separately searched the 454 and Solexa data (*SI Text*). With the default cut-off we recovered miRNAs with high sensitivity and specificity (Fig. 5 *A* and *B*). miRDeep identified 70 novel potential miRNAs, which were further curated (see *SI Text*). We thus report a subset of 61 high-confidence miRNAs (Table S6).

We subjected 20 miRNA candidates to Northern blot analysis and successfully validated 13 of them (see Fig. 5*C*). Candidates not observed by Northern blotting may be below detection threshold. In support of this, a more sensitive Taqman assay was used to validate 11 of 11 novel candidates tested (Fig. 3), 4 of which had also been validated by Northern blot analysis (Fig. 5*C*). In total, 20 novel candidates were validated.

The phylogenetic analysis of planarian miRNAs may be particularly informative as planarians are an outgroup relative to animal model systems used by the majority of researchers. miRNAs can be grouped into families based on sequence similarity at their 5′ end (7). Our novel miRNAs increase the number of planarian miRNA families from 37 to 79 (Fig. S3). The planarian miRNAs share 22 families with mammals and 33 with flies and with nematodes. Thus, we find planaria to resemble Ecdysozoa (flies, nematodes) more

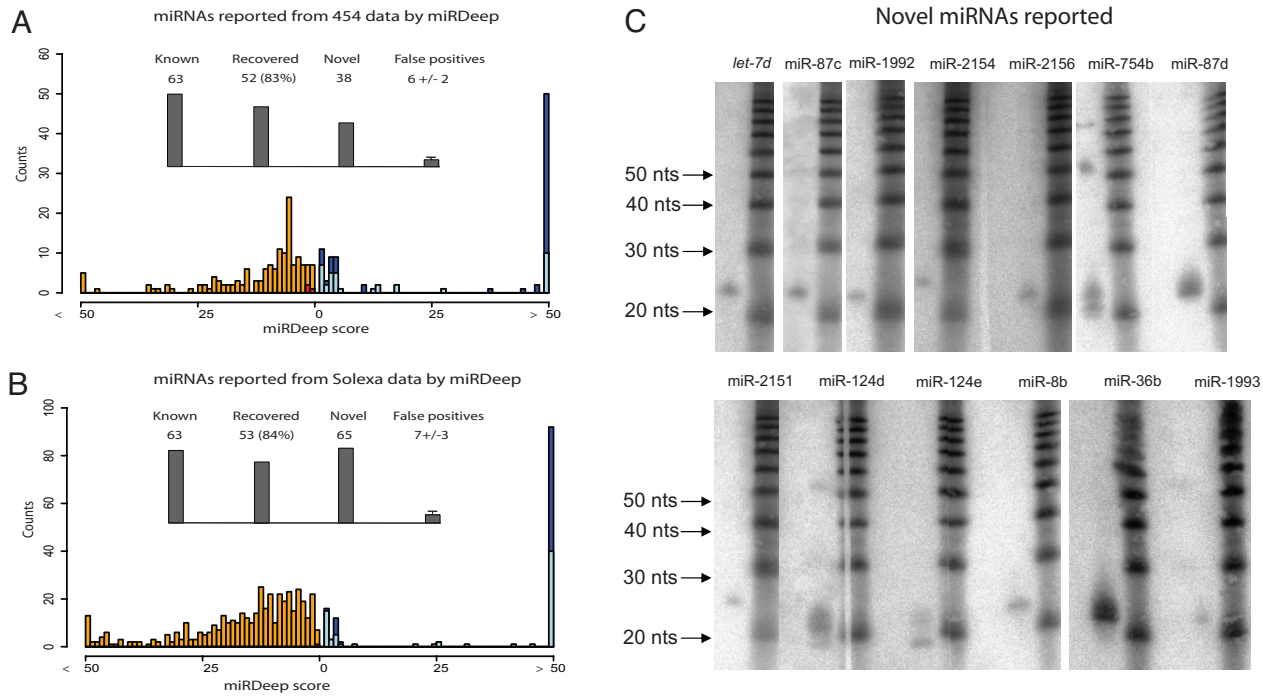


Fig. 5. Novel planarian miRNAs. (*A*) miRNAs reported from 454 data by miRDeep. miRDeep scores are shown as a color-encoded histogram (score cut-off: 1). Known miRNAs, dark blue; novel miRNAs, light blue. Known miRNAs below the cut-off are plotted in red (false negatives). The number of false positives was estimated by miRDeep (20). (*B*) miRNAs reported from Solexa data by miRDeep; legend same as *A*. (*C*) Validation of novel miRNAs. 13 miRNAs were validated by Northern blot analysis, 11 by qPCR (Fig. 3). In some cases, miRNA precursors are also detected.

than mammals from the miRNA phylogeny. Interestingly, the majority (45 of 79) of planarian miRNA families do not show sequence similarity to known miRNAs. The presence of the miR-1992 family (21) combined with the absence of the miR-1994 family gives evidence to the hypothesis that flatworms are the sister group to the other lophotrochozoans.

More than a Dozen miRNAs Are Likely Linked to Neoblast Biology. To identify miRNAs up-regulated in neoblasts, we calculated the miRNA expression fold-change between the untreated whole worm and isolated neoblast deep-sequencing datasets (*SI Text* and Table S7). 13 miRNAs were up-regulated by >2-fold in the neoblast sample. As an independent control, we used qPCR to profile the expression of these miRNAs in filtered cells enriched in neoblasts vs. untreated planarians. These were all up-regulated by >30% in the isolated neoblasts, whereas a number of other profiled miRNAs did not change.

To rule out that miRNA up-regulation may have been caused by cell dissociation or cell sorting, we used qPCR in independently obtained samples to calculate miRNA expression fold-changes between untreated and irradiated animals. We found 10 of 13 of the miRNAs of interest to be >25% down-regulated in the irradiated sample (see Table 2 and Fig. 3). These data suggest that a small subset of miRNAs is significantly up-regulated in neoblasts. Notably, miRNA genes comprised in clusters (miR-2d, miR-13, miR-71b, miR-752 and *let-7b*, miR-36b) as well as the miRNAs belonging to the same families (*let-7*, miR-2/miR-13) show a similar differential expression (Fig. 3).

Interestingly, most of the up-regulated miRNAs in neoblasts belong to conserved families. The *let-7* family has previously been associated with stem-cell identity. However, previous studies indicate that *let-7* is down-regulated posttranscriptionally in stem cells (22). In flies, miR-2 and miR-13 target the proapoptotic genes *grim*, *reaper*, and *sickle* (23). We find that all 4 miRNAs in genomic cluster containing the planarian miR-2 and miR-13 are up-regulated in neoblasts, suggesting that these miRNAs are important for neoblast maintenance or neoblast-related function. Additionally, miRNAs that are typically expressed in specific somatic tissues such as miR-124 (brain tissues) and miR-1 and miR-133 (muscle tissues) were down-regulated in neoblasts.

Discussion

By using massive quantitative deep sequencing, we have annotated small RNA species in *S. mediterranea*. We have doubled the number of planarian miRNAs, and have validated a large fraction of our novel miRNAs. We find that the small RNA-length peak at nucleotide 22 disappears completely when miRNAs are removed (Fig. 2), suggesting that a diminishing number of miRNAs or other Dicer products remain to be discovered in *S. mediterranea*. We were also unable to detect evidence for phased processing of longer transcripts by Dicer. Moreover, we annotated more than one-million unique piRNA sequences that locate to genomic clusters. piRNAs have previously been well-described in Deuterostomes and ecdysozoans (15, 16, 24–26). We report ≈1.2 million unique planarian

Table 2. Ten miRNAs up-regulated in neoblasts

miRNAs	Fold change neo/untr (454)	Fold change untr/irr (qPCR)	Fold change untr/irr (Solexa)
let-7a	2.4	3.3	5.1
let-7b, miR-36b	2.1, 3.8	2.5, 2.0	2.1, 0.9
miR-2a	3.7	1.4	1.0
miR-2d, miR-13, miR-71b, miR-752	3.4, 4.9, 3.0, 7.0	2.0, 3.3, 2.5, 6.5	1.9, 5.5, 2.3, 6.2
miR-756	2.0	1.4	1.6
miR-2160	5.0	2.0	1.9

miRNAs listed in the same field locate to the same genomic cluster. Neo, neoblast sample; untr, untreated sample; irr, irradiated sample.

piRNAs locating to more than 100 genomic clusters, and thus give a first comprehensive description of piRNAs in lophotrochozoans. We find that piRNA features characteristic for piRNA biogenesis (sequence biases, 10-nt overlap) are shared in all 3 metazoan superphylae. Planarian piRNAs also share specific characteristics with either mammalian or fly piRNAs. Planarian primary piRNAs, like those in the fly, tend to map antisense to transposable elements, suggesting that planaria may defend their genome against transposons similarly to flies. However, from a different point of view, planarian piRNA biology resembles that of the mouse more than the fly. First, flatworm piRNAs associate with transposons as much as mouse prepachytene piRNAs. Second, the expression of planarian piRNAs is dispersed between numerous clusters, similar to what has recently been observed in the mouse (17). Third, we find no planarian clusters containing many transposon fragments akin to the characteristic fly master loci. Together, our data indicate that the piRNA pathway has undergone complex evolution.

We find that at least 10 miRNAs are up-regulated in neoblast samples. Deep sequencing and qPCR controls show that these are down-regulated in the irradiated samples depleted of neoblasts, indicating that they may be specific to neoblast biology. These miRNAs include all 4 miRNAs from a genomic cluster that contains miR-2 and miR-13, miRNAs known to inhibit proapoptotic genes in fly (23). We also find that at least 2 members (*let-7a* and *let-7b*) of the highly conserved *let-7* family are up-regulated in neoblasts. Up-regulation of *let-7* in neoblast samples was paralleled by *let-7* down-regulation in irradiated samples. These findings are surprising because *let-7* and its family members are known to be depleted in mammalian stem cells (22, 27) and have been shown in numerous species to repress cell proliferation and promote differentiation (reviewed in ref. 28). However, recent studies have shown that cells with the morphological appearance of neoblasts can be resolved into subtypes, and planarian stem cells may maintain proliferative activity after commitment (4, 29). Thus, high *let-7* levels in neoblast samples may be derived from neoblasts that are exiting the stem-cell state and committing to a differentiation lineage. Further *let-7* expression analyses, therefore, may help elucidate the specification of neoblast lineages.

Although planarian stem cells are collectively totipotent because they can give rise to both the somatic and germ lineages in the adult, our analyses indicate that planarians harbor only 1 miRNA (miR-92) known to be highly expressed in mammalian embryonic stem cells (30). However, we found no evidence that its 2 family members are up-regulated in neoblasts. Taken together, expression of miRNAs in planarian neoblasts share little if any similarity with mammalian embryonic stem cells, which may reflect both the adult nature of planarian stem cells as well as the inherent in vitro versus in vivo differences between these 2 populations of animal stem cells.

Finally, the small RNA profile of neoblasts resemble mouse and fly germ-line stem cells in being dominated by piRNAs. Because the genomic contents of germ-line cells and neoblasts are potentially immortal, both cell types need to strictly control their genome integrity during transmission to future generations, and particularly, to protect it against the uncontrolled propagation of mobile genetic elements. piRNAs have been shown selectively to silence transposons in the fly and mouse genomes (reviewed in ref. 18) and it is likely that piRNAs play such a role in planaria. Further studies are needed to determine whether planarian piRNAs also play a critical role in epigenetic silencing through DNA/chromatin methylation like their germ-line homologs (17).

Methods

Sample Preparation and Sequencing. Planarians from the clonal, asexual CIW4 strain of *S. mediterranea* were starved for 1 week before all experiments. Planarian total RNA was isolated by using TRIzol (Invitrogen). Planarians for the irradiated samples were exposed to 60 Gy and RNA was extracted 8 days after irradiation. FACS sorting was performed as described in *SI Text*. Solexa and 454 sequencing was performed by using the manufacturer's protocol.

miRNA Identification. Detection of novel miRNAs was performed as in ref. 20. For more details, see *SI Text*.

Northern Blot Analysis and qPCR. Validation of miRNA and piRNA candidates was performed by Northern blot analysis (Table S8) as described previously (31). qPCR (Taqman miRNA custom assays, ABI) was used to quantify the expression fold-change of 35 miRNAs. cDNA was synthesized from 50 ng of total RNA from either irradiated or untreated animals. Samples without reverse transcriptase served as a negative control template. Each measurement was performed in triplicate. Two biological replicates were used. Threshold cycle values are relative to expression detected for the ubiquitously expressed control mRNA *ura4* (*SI Text*). *Relative expression of miRNAs is given as log2 of* $2^{-\Delta\Delta Ct}$ values.

ACKNOWLEDGMENTS. We thank Astrid Ferlinz from Applied Biosystems for the custom ABI Taqman primers. Sam Griffith-Jones facilitated the manuscript through a rapid assignment of miRBase names to the novel planarian miRNAs. Erik A. Sperling and Kevin J. Peterson gave valuable advice on miRNA phylogeny.

1. Reddien PW, Sanchez Alvarado A (2004) Fundamentals of planarian regeneration. *Annu Rev Cell Dev Biol* 20:725–757.
2. Randolph H (1892) The regeneration of the tail in lumbriculus. *J Morphol* 7:317–344.
3. Robb SM, Ross E, Sanchez Alvarado A (2008) SmedGD: *The Schmidtea mediterranea* genome database. *Nucleic Acids Res* 36:D599–D606.
4. Eisenhoffer GT, Kang H, Sanchez Alvarado A (2008) Molecular analysis of stem cells and their descendants during cell turnover and regeneration in the planarian *Schmidtea mediterranea*. *Cell Stem Cell* 3:327–339.
5. Reddien PW, Bermange AL, Murfitt KJ, Jennings JR, Sanchez Alvarado A (2005) Identification of genes needed for regeneration, stem cell function, and tissue homeostasis by systematic gene perturbation in planaria. *Dev Cell* 8:635–649.
6. Ghildiyal M, Zamore PD (2009) Small silencing RNAs: An expanding universe. *Nat Rev Genet* 10:94–108.
7. Bartel DP (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* 116:281–297.
8. Yin VP, et al. (2008) Fgf-dependent depletion of microRNA-133 promotes appendage regeneration in zebrafish. *Genes Dev* 22:728–733.
9. Reddien PW, Oviedo NJ, Jennings JR, Jenkin JC, Sanchez Alvarado A (2005) SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science* 310:1327–1330.
10. Palakodeti D, Smielewska M, Lu YC, Yeo GW, Graveley BR (2008) The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. *RNA* 14:1174–1186.
11. Palakodeti D, Smielewska M, Graveley BR (2006) MicroRNAs from the planarian *Schmidtea mediterranea*: A model system for stem cell biology. *RNA* 12:1640–1649.
12. Gonzalez-Estevez C, Arseni V, Thambyrajah RS, Felix DA, Aboobaker AA (2009) Diverse miRNA spatial expression patterns suggest important roles in homeostasis and regeneration in planarians. *Int J Dev Biol* 53:493–505.
13. Winter J, Jung S, Keller S, Gregory RI, Diederichs S (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat Cell Biol* 11:228–234.
14. Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* 318:761–764.
15. Brennecke J, et al. (2007) Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* 128:1089–1103.
16. Gunawardane LS, et al. (2007) A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila*. *Science* 315(5818):1587–1590.
17. Aravin AA, et al. (2008) A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* 31:785–799.
18. O'Donnell KA, Boeke JD (2007) Mighty Piwis defend the germline against genome intruders. *Cell* 129:37–44.
19. Garcia-Fernandez J, et al. (1995) High copy number of highly similar mariner-like transposons in planarian (Platyhelminthe): Evidence for a trans-phyla horizontal transfer. *Mol Biol Evol* 12:421–431.
20. Friedlander MR, et al. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 26:407–415.
21. Wheeler BM, et al. (2009) The deep evolution of metazoan microRNAs. *Evol Dev* 11:50–68.
22. Thomson JM, et al. (2006) Extensive post-transcriptional regulation of microRNAs and its implications for cancer. *Genes Dev* 20:2202–2207.
23. Stark A, Brennecke J, Russell RB, Cohen SM (2003) Identification of *Drosophila* microRNA targets. *PLoS Biol* 1:E60.
24. Aravin A, et al. (2006) A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* 442:203–207.
25. Girard A, Sachidanandam R, Hannon GJ, Carmell MA (2006) A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442:199–202.
26. Grivna ST, Beyret E, Wang Z, Lin H (2006) A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* 20:1709–1714.
27. Viswanathan SR, Daley GQ, Gregory RI (2008) Selective blockade of microRNA processing by Lin28. *Science* 320:97–100.
28. Roush S, Slack FJ (2008) The let-7 family of microRNAs. *Trends Cell Biol* 18:505–516.
29. Higuchi S, et al. (2007) Characterization and categorization of fluorescence activated cell sorted planarian stem cells by ultrastructural analysis. *Dev Growth Differ* 49:571–581.
30. Calabrese JM, Seila AC, Yeo GW, Sharp PA (2007) RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc Natl Acad Sci USA* 104:18097–18102.
31. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. *Science* 294:853–858.

Molecular Cell
Article

A Human snoRNA with MicroRNA-Like Functions

Christine Ender,[1,6] Azra Krek,[2,3,6] Marc R. Friedländer,[2] Michaela Beitzinger,[1] Lasse Weinmann,[1] Wei Chen,[4] Sébastien Pfeffer,[5] Nikolaus Rajewsky,[2,*] and Gunter Meister[1,*]
[1]Center for Integrated Protein Science Munich (CIPSM), Laboratory of RNA Biology, Max Planck Institute of Biochemistry, Am Klopferspitz 18, 82152 Martinsried, Germany
[2]Max Delbrück Center for Molecular Medicine, Robert-Rössle-Strasse 10, 13125 Berlin, Germany
[3]Department of Physics, New York University, 4 Washington Place, New York, NY 10003, USA
[4]Department of Human Molecular Genetics, Max Planck Institute of Molecular Genetics, Ihnestrasse 73, 14195 Berlin, Germany
[5]IBMP-CNRS, 12 Rue du General Zimmer, 67084 Strasbourg Cedex, France
[6]These authors contributed equally to this work
*Correspondence: rajewsky@mdc-berlin.de (N.R.), meister@biochem.mpg.de (G.M.)
DOI 10.1016/j.molcel.2008.10.017

SUMMARY

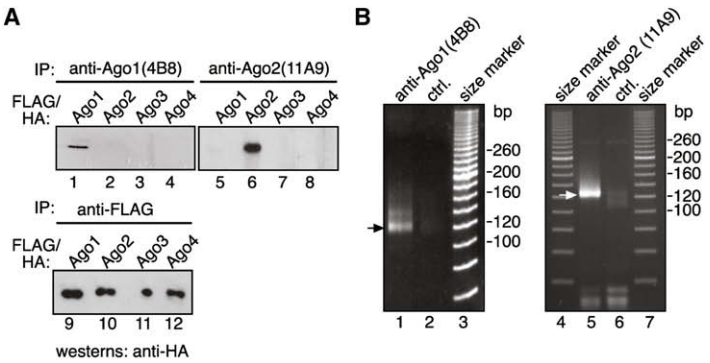
Small noncoding RNAs function in concert with Argonaute (Ago) proteins to regulate gene expression at the level of transcription, mRNA stability, or translation. Ago proteins bind small RNAs and form the core of silencing complexes. Here, we report the analysis of small RNAs associated with human Ago1 and Ago2 revealed by immunoprecipitation and deep sequencing. Among the reads, we find small RNAs originating from the small nucleolar RNA (snoRNA) ACA45. Moreover, processing of ACA45 requires Dicer activity but is independent of Drosha/DGCR8. Using bioinformatic prediction algorithms and luciferase reporter assays, we uncover the mediator subunit CDC2L6 as one potential mRNA target of ACA45 small RNAs, suggesting a role for ACA45-processing products in posttranscriptional gene silencing. We further identify a number of human snoRNAs with microRNA (miRNA)-like processing signatures. We have, therefore, identified a class of small RNAs in human cells that originate from snoRNAs and can function like miRNAs.

INTRODUCTION

Small noncoding RNAs, including microRNAs (miRNAs), short interfering RNAs (siRNAs), and Piwi-interacting RNAs (piRNAs), are important regulators of gene expression (Filipowicz et al., 2005; Meister and Tuschl, 2004; Seto et al., 2007). miRNAs and siRNAs guide sequence-specific cleavage, deadenylation, or translational repression of target mRNAs (Chen and Rajewsky, 2007; Pillai et al., 2007). piRNAs are specifically expressed in testes (Seto et al., 2007) and control retrotransposition in the mammalian germ line (Aravin et al., 2007).

In many gene-silencing pathways, small RNAs are generated from double-stranded RNA (dsRNA) molecules by distinct processing steps (Tomari and Zamore, 2005). miRNA genes are transcribed by RNA polymerases II or III as primary miRNAs that are further processed to hairpin-structured miRNA precursors (pre-miRNAs) by the nuclear microprocessor complex containing the RNase III enzyme Drosha and its cofactor DGCR8 (Borchert et al., 2006; Denli et al., 2004; Gregory et al., 2004; Landthaler et al., 2004; Lee et al., 2003, 2004). Pre-miRNAs are transported to the cytoplasm, where the RNase III enzyme Dicer cleaves off the loop of the miRNA hairpin, thereby generating a short dsRNA of about 20–25 nucleotides (nt) in length (Bohnsack et al., 2004; Grishok et al., 2001; Hutvágner et al., 2001; Lund et al., 2004). Such dsRNA intermediates are subsequently unwound, and the single-stranded mature miRNA is incorporated into effector complexes often referred to as miRNPs (Mourelatos et al., 2002). In the siRNA pathway or RNA interference (RNAi), long dsRNA is processed by Dicer as well (Bernstein et al., 2001). The mature siRNA is incorporated into the RNA-induced silencing complex (RISC). The biogenesis of piRNAs is only poorly understood and probably does not require the function of Drosha or Dicer.

Argonaute (Ago) proteins are the cellular binding partners of small RNAs and form the core of gene silencing effector complexes (Parker and Barford, 2006; Peters and Meister, 2007). In humans, eight different Argonaute genes exist, which can be phylogenetically divided into four Ago and four Piwi subfamily members (Peters and Meister, 2007; Tolia and Joshua-Tor, 2007). Whereas Piwi proteins interact with piRNAs in the germ line (Seto et al., 2007), Ago subfamily members associate with miRNAs in somatic cells. Argonaute proteins are generally characterized by Piwi-Argonaute-Zwille (PAZ) and PIWI domains (Parker and Barford, 2006; Peters and Meister, 2007). A third domain, termed MID domain, anchors the 5′ end of the small RNA (Ma et al., 2005; Parker et al., 2005). The PAZ domain binds the 3′ end of the small RNA, and the PIWI domain, which is structurally similar to RNase H, cleaves the complementary target RNA (Parker and Barford, 2006; Patel et al., 2006; Tolia and Joshua-Tor, 2007). However, not all Argonaute proteins are endonucleases, although critical residues within the PIWI domain are conserved. In mammals, only Ago2 has been shown to act as endonuclease in RNAi (Liu et al., 2004; Meister et al., 2004). Argonaute proteins with endonuclease activity are often referred to as Slicers. Although Ago subfamily members have been extensively studied in the past, only little is known about their individual small RNA-binding specificities. It has been reported that all Ago proteins bind miRNAs or siRNAs indiscriminately of their



C

	Ago1	*Ago2*
Total # of reads	**20448**	**42604**
Have known annotation	17803	31973
# miRNAs	**16628**	**31212**
# sno/scaRNAs	**243**	**43**
- CDBox	178	11
- HAcaBox	32	22
- scaRNA	33	10
# mRNAs	**285**	**358**
- 3' UTR	99	178
- 5' UTR	52	46
- Coding exon	134	134
# misc ncRNAs	**484**	**130**
- rRNA	313	52
- tRNA	159	71
- scRNA	2	3
- snRNA	10	4
# Transposable elements (LINE, SINE, LTR)	**157**	**92**
# Other	**6**	**38**

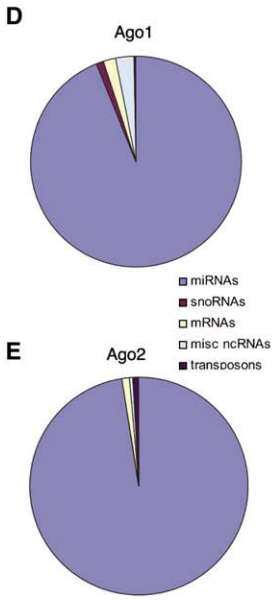


Figure 1. Small RNAs Associated with Ago1 and Ago2 Complexes

(A) Characterization of monoclonal anti-Ago1 and anti-Ago2 antibodies. FLAG/HA-tagged Ago1 through 4 were subjected to immunoprecipitations using anti-Ago1(4B8) (lanes 1–4), anti-Ago2(11A9) (lanes 5–8), and anti-FLAG (lanes 9–12). Immunoprecipitated FLAG/HA-Ago proteins were analyzed by western blotting using anti-HA antibodies.

(B) Endogenous Ago1 (lane 1) and Ago2 (lane 5) complexes were immunoprecipitated using the specific monoclonal antibodies described in (A). An anti-FLAG (lane 2) or an anti-GST antibody (lane 6) served as controls. Coimmunoprecipitated RNAs were extracted, cloned, and sequenced. Cloned PCR products containing 5′ adaptors, poly(A) tails, and 3′ adaptor sequences were loaded onto an agarose gel and visualized by ethidium bromide staining.

(C) Summary of the sequencing data obtained from deep sequencing of human Ago1 and Ago2 associated small RNAs.

(D and E) Schematic representation of individual small RNA classes that are associated with human Ago1 or Ago2 complexes.

sequence (Liu et al., 2004; Meister et al., 2004). However, a recent study analyzed small RNAs that are associated with human Ago2 and Ago3 and suggested that Ago proteins might have preferences for individual miRNA species, although all miRNAs that have been investigated bind to both Ago2 and Ago3 (Azuma-Mukai et al., 2008).

Here, we report the characterization of small RNAs associated with human Ago1 and Ago2 by immunoprecipitation and deep sequencing. We find that Ago1 and Ago2 bind to similar sets of miRNAs, although some miRNAs are more prominent in Ago2 libraries and vice versa. More importantly, we find small RNAs that originate from small nucleolar RNAs (snoRNAs). snoRNAs are nucleolar noncoding RNAs, which have important functions in the maturation of other noncoding RNAs such as ribosomal RNAs (rRNAs) or small nuclear RNAs (snRNAs) (Matera et al., 2007). We demonstrate that the bona fide snoRNA ACA45 is processed to small 20- to 25-nt-long RNAs that stably associate with Ago proteins. Processing is independent of the Drosha/DGCR8 complex but requires Dicer. Finally, we identify a cellular target mRNA that is regulated by the ACA45-derived small RNA, indicating that snoRNA-derived small RNAs can function like miRNAs.

RESULTS

Small RNAs Associated with Human Ago1 and Ago2

Different Ago proteins associate with the same miRNA species regardless of their sequence (Azuma-Mukai et al., 2008; Meister et al., 2004). However, the spectrum of Ago-associated small RNAs in human somatic cells is presently not known. Therefore, we used monoclonal antibodies specific to human Ago1 (Ago1 [4B8]) (Beitzinger et al., 2007) and Ago2 (Ago2 [11A9]) (Rudel et al., 2008) for Ago isolation from total HEK293 cell lysates (Figure 1A). Coimmunoprecipitated RNAs were extracted and cloned without size fractionation (Figure 1B). Using 454 deep sequencing, we obtained 20448 reads from the Ago1-associated and 42604 reads from the Ago2-associated small RNA libraries (Figures 1C–1E). Using a Dicer substrate identification algorithm (Friedländer et al., 2008), the presence of 166 known miRNAs in the combined Ago1 and Ago2 libraries was confirmed. We next investigated whether miRNAs are differentially bound to Ago1 or Ago2 in HEK293 cells (Table S1 available online). All miRNAs that are present in the libraries bind to Ago1 as well as Ago2. Similarly to the published data on Ago2 and Ago3 miRNA association (Azuma-Mukai et al., 2008), some miRNAs are more highly represented in one or the other library, suggesting a preferential Ago binding.

Processing of Functional Small RNAs from the Bona Fide snoRNA ACA45

In the Ago-associated RNA libraries, we have identified small RNAs with a length of about 20–22 nt that originate from snoRNAs particularly from ACA45 (Figure 2A). Notably, the sequenced reads derive only from the hairpin formed by the 3′ half of ACA45. The found reads are conserved in mammals (Figure 2B), suggesting that they are, indeed, specific processing products.

Although ACA45 was identified in a screen for functional snoRNAs (Kiss et al., 2004), it is conceivable that it represents a miRNA gene that has been misannotated as snoRNA. Due to

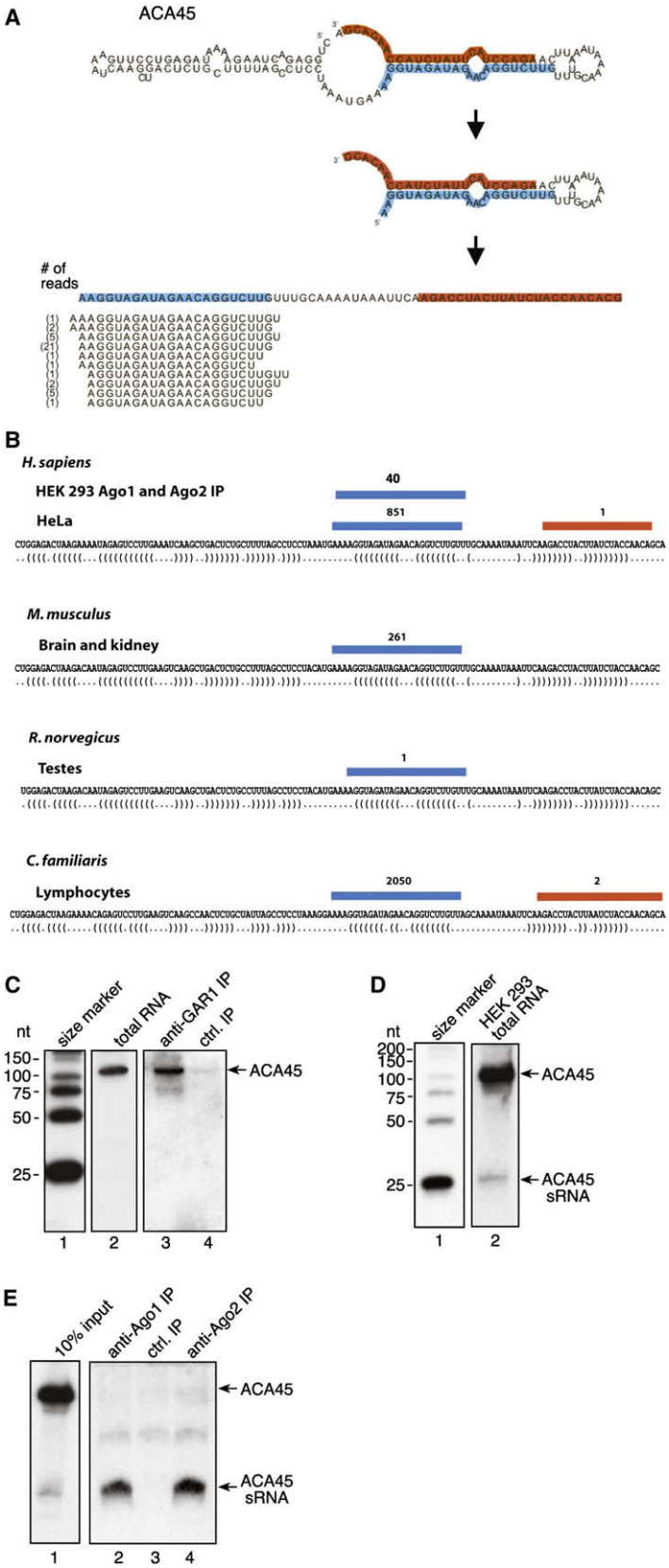


Figure 2. snoRNA ACA45 Is Processed to Small RNAs

(A) Sequence reads found in Ago1 and/or Ago2 libraries are highlighted in orange and blue. Numbers of individual sequences are shown in parentheses.

(B) Specific ACA45-processing products are conserved across species. The ACA45 sequences for *Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, and *Canis familiaris* are shown by sequence and structure. Independent deep sequencing data from each species have been obtained and mapped against the corresponding homolog (data sets referenced in Experimental Procedures). The bars indicate the positions of mapped reads; consensus positions are shown where small variations exist. The blue bars represent putative "mature" Dicer products, while the red bars represent putative "star" products. The numbers above the bars indicate the number of reads.

(C) ACA45 associates with the snoRNP component GAR-1. HEK293 lysates were immunoprecipitated using anti-GAR-1 (lane 3) or control antibodies (lane 4). Coimmunoprecipitated RNA was extracted and analyzed by northern blotting using a probe specific to ACA45. Lane 2 shows total RNA extracted from 10% of the input lysate, and lane 1 shows a size marker.

(D) Total RNA from HEK293 cells (30 μg) was blotted, and the membrane was incubated with a probe complementary to nt 65–85 (lane 2) of ACA45. A size marker is shown in lane 1.

(E) Endogenous Ago1 (lane 2) or Ago2 (lane 4) were immunoprecipitated from HEK293 lysates, and the coimmunoprecipitated RNAs were extracted and analyzed by northern blotting using the probe described in (D). Lane 1 shows 10% of the extracts used for the immunoprecipitations, and lane 3 shows a control immunoprecipitation.

their specific structures and functions, snoRNAs can be grouped in H/ACA and Box C/D class snoRNAs. snoRNAs associate with specific protein components such as GAR-1 (H/ACA) or fibrillarin (Box C/D) to form functional snoRNPs (Matera et al., 2007). In order to prove that ACA45 is, indeed, a functional snoRNA, we analyzed GAR-1 binding to ACA45 (Figure 2C). Endogenous GAR-1 was immunoprecipitated using anti-GAR-1 antibodies. Associated RNAs were extracted and further analyzed by northern blotting using the probe specific to ACA45. Indeed, full-length ACA45 was readily detectable in the anti-GAR-1 (lane 2), but not in control immunoprecipitates (lane 3). Our data, therefore, confirm that ACA45 represents a functional snoRNA.

We next validated the processing of ACA45 to small RNAs by northern blotting (Figure 2D). A probe complementary to the 5′ arm (Figure 2A, indicated in blue) detected the full-length ACA45 snoRNA as well as a band of ~22–23 nt in total RNA, indicating that a portion of the cellular ACA45 pool is, indeed, processed to small RNAs. Using quantitative northern blotting, we analyzed ACA45 sRNA molecule numbers per cell (data not shown). We find that less than 1000 molecules are present per cell, which is similar to a low abundant miRNA (Lim et al., 2003). Since only a minor portion of ACA45 is processed to small RNAs, we next investigated whether ACA45 processing products are specifically enriched in Ago protein complexes (Figure 2E). Endogenous Ago1 (lane 2) or Ago2 (lane 4) were immunoprecipitated, and associated RNAs were analyzed by northern blotting against ACA45-processing products. Consistent with the cloning data, the small RNA derived from ACA45 was enriched in Ago1 as well as Ago2 immunoprecipitates, indicating that ACA45-processing products specifically associate with Ago proteins. Therefore, we refer to this functional small RNA as ACA45 small RNA (ACA45 sRNA).

ACA45 Small RNAs Can Function Like miRNAs

The striking similarity of ACA45-processing products to miRNA precursors prompted us to investigate whether ACA45 sRNAs are functionally similar to miRNAs. We generated a luciferase reporter construct containing a complementary binding site for the abundant 5′ arm of the snoRNA precursor (Figure 3A). Luciferase activity was strongly increased when the endogenous ACA45-derived small RNAs were inhibited using 2′-O-methylated (2′-OMe) antisense inhibitors (Figure 3A). Moreover, luciferase activity was also increased when the RNAi endonuclease Ago2 was depleted (Figure 3B), indicating that small RNAs that are processed from ACA45 can function like miRNAs.

ACA45 Processing Is Independent of the Drosha/DGCR8 Complex but Requires Dicer

The cleavage signature of the stem-loop-structured processing intermediate is different than the typical 2 nt 3′ overhangs generated by Drosha. Therefore, we analyzed whether ACA45 processing requires activity of the Drosha/DGCR8 complex using in vitro as well as in vivo approaches (Figures 3B and 3C). FLAG/HA(FH)-tagged DGCR8 was immunoprecipitated, and the immunoprecipitate was incubated with either a ^{32}P-labeled primary miR-27a transcript or ACA45. A specific cleavage product representing pre-miR-27a was observed in the anti-DGCR8 immunoprecipitates, whereas no signal was observed when ACA45 was used as substrate. We further investigated Drosha requirements using the luciferase reporter construct described above (Figure 3B). Indeed, we did not observe elevated luciferase activity upon Drosha depletion (siRNAs have been validated in Landthaler et al. [2004]), whereas luciferase activity of a miR-19b-responsive reporter was significantly increased. Taken together, our results suggest that ACA45 processing is independent of the Drosha/DGCR8 complex.

Next, we investigated Dicer requirements for ACA45 processing. It has been demonstrated that Ago proteins form a stable complex with Dicer, and Dicer activity can be coimmunoprecipitated with antibodies against Ago proteins (Gregory et al., 2005; Maniataki and Mourelatos, 2005; Meister et al., 2005). Therefore, FH-tagged Ago proteins, as well as FH-Dicer, was immunoprecipitated from HEK293 lysates and incubated with ^{32}P-labeled pre-miR-27a or full-length ACA45 (Figure 3D). As expected, both FH-Ago2 and FH-Dicer immunoprecipitates efficiently processed the miR-27a precursor (Figure 3D, left panel). Furthermore, FH-Ago1, FH-Ago2, and FH-Dicer immunoprecipitates processed the ^{32}P-labeled full-length ACA45 as well (Figure 3D, right panel), suggesting that Dicer is required for the generation of ACA45 small RNAs. To further investigate Dicer's function in ACA45 processing, we analyzed whether Dicer alone is sufficient for ACA45 processing in vitro. ^{32}P-labeled ACA45 was incubated with increasing amounts of recombinant Dicer, and cleavage products were analyzed by RNA-PAGE (Figure 3E). Indeed, recombinant Dicer produced small RNAs from the full-length ACA45 in a concentration-dependent manner, suggesting that Dicer alone is sufficient for ACA45 processing. Notably, Dicer generates longer RNAs as well, which might represent processing intermediates (see asterisk in Figure 3D). Finally, we analyzed the role of Dicer in ACA45 processing in vivo. Total RNA from mouse embryonic stem (ES) cells carrying homozygous or heterozygous Dicer deletions (Murchison et al., 2005) was analyzed for the presence of ACA45 small RNAs by semiquantitative real-time PCR (qRT-PCR) (Figure 3F). Strikingly, no PCR product was detectable in the Dicer$^{-/-}$ cells, whereas a PCR product originating from the ACA45 small RNA was readily detectable in Dicer$^{+/-}$ cells. Notably, the full-length ACA45 was present in both Dicer$^{-/-}$ and Dicer$^{+/-}$ cells. Similar results were obtained when total RNA from Dicer$^{-/-}$ and Dicer$^{+/-}$ cells was analyzed by northern blotting using a probe complementary to the ACA45 small RNA (Figure 3G). In summary, our data indicate that Dicer processes ACA45 to small RNAs independently of the Drosha-containing microprocessor complex.

Validation of an Endogenous ACA45-Derived Small RNA Target

It is thought that complementary Watson-Crick base pairing of the seed (nucleotides 2–8 counted from the 5′ end) is a key feature of miRNA:mRNA target recognition. It is also known that highly conserved 7-mers in 3′UTRs are often complementary to seed sequences of known miRNAs (Chen and Rajewsky, 2007). Remarkably, the seed of ACA45 22-nt-long processing product is perfectly complementary to a significantly conserved 3′UTR motif (top 3% of all possible seed sites). Using the miRNA target prediction algorithm PicTar (Krek et al., 2005), we have predicted target mRNAs for the ACA45-derived small RNA

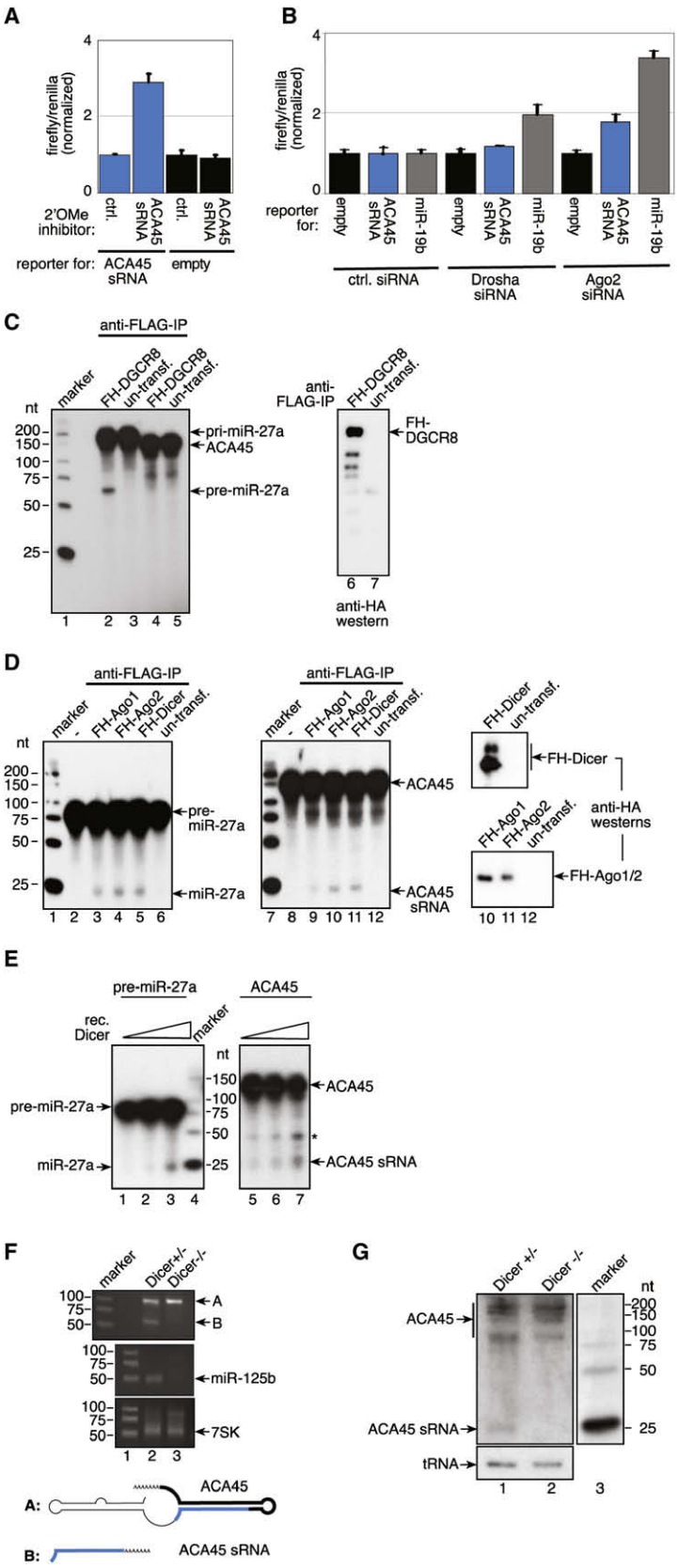


Figure 3. ACA45 Processing Requires Dicer but Is Independent of Drosha

(A) A luciferase reporter construct containing a perfectly complementary binding site for the ACA45 sRNA or the empty vector was cotransfected with 2′-O-methylated antisense inhibitors directed against the ACA45 sRNA.

(B) The luciferase reporter described in (A), the empty vector, and a luciferase reporter containing a complementary binding site to miR-19b were transfected into HEK293 cells that have been pretransfected with control siRNAs, siRNAs directed against Drosha, and siRNAs against Ago2. Firefly luciferase activity was normalized to Renilla activity. Error bars are derived from four individual experiments.

(C) FH-DGCR8 or untreated cells were immunoprecipitated using anti-FLAG antibodies. Immunoprecipitates were incubated with ^{32}P-labeled pri-miRNA-27a (lanes 2 and 3) or ACA45 (lanes 4 and 5). Lane 1 represents a size marker, and lanes 6 and 7 represent the protein input.

(D) FH-Ago2 (lanes 4 and 10), FH-Ago1 (lanes 3 and 9), and FH-Dicer (lanes 5 and 11) were incubated with ^{32}P-labeled pre-miR-27a (lanes 2–6) or ACA45 (lanes 8–12) and analyzed by RNA PAGE. In lanes 6 and 12, lysate from untransfected HEK293 cells was used. Lanes 13–15 show anti-HA western blots of the protein input. Lanes 1 and 7 show size markers.

(E) ^{32}P-labeled pre-miR-27a (lanes 1–3) or ACA45 (lanes 5–7) were incubated with increasing amounts of recombinant Dicer. Cleavage products were analyzed by RNA PAGE. Lane 4 shows a size marker. A putative processing intermediate is indicated by an asterisk.

(F) Total RNA from Dicer$^{-/+}$ (lane 2) or Dicer$^{-/-}$ cells was analyzed by semi-qRT-PCR using primers specific for the ACA45 sRNA (upper panel), miR-125b (middle panel), and 7SK RNA (lower panel). The origin of the PCR products indicated as A and B are highlighted in bold below the figure.

(G) Total RNA from Dicer$^{+/-}$ (lane 1) or Dicer$^{-/-}$ (lane 2) cells was analyzed by northern blotting using probes specific for the ACA45 small RNA described above. Lane 3 shows a size marker.

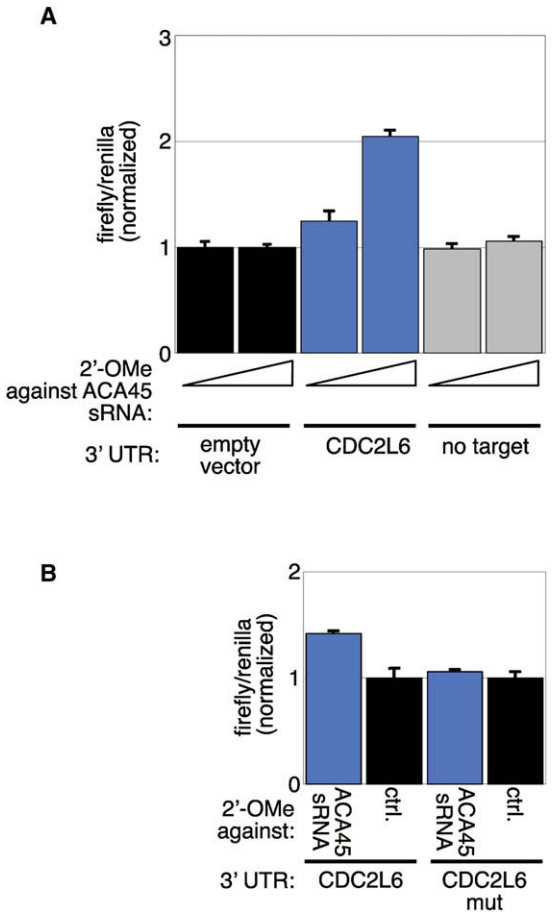


Figure 4. ACA45-Derived Small RNAs Regulated CDC2L6 Expression

(A) Luciferase reporter constructs containing the 3′UTR of CDC2L6 or BAP-1 (no target) or the empty vector were cotransfected with increasing concentrations of 2′-OMe inhibitors against the ACA45 sRNA. Firefly luciferase activity was normalized to Renilla activity. Error bars are derived from four individual experiments.

(B) Luciferase reporter constructs containing the 3′UTR of CDC2L6 or the CDC2L6 3′UTR with mutated ACA45 sRNA-binding sites were cotransfected with 2′-OMe inhibitors against the ACA45 sRNA. Firefly luciferase activity was normalized to Renilla activity. Error bars are derived from four individual experiments.

(data not shown). For experimental validation, we fused a number of 3′UTRs that we selected from the predicted target mRNAs to a luciferase reporter gene. Luciferase reporter constructs were cotransfected with 2′-OMe oligonucleotides antisense to the ACA45 small RNA. Many of the tested 3′UTRs, however, did not respond to the 2′-OMe inhibitors, suggesting that the small RNA does not target these mRNAs or that small RNA-target mRNA interactions are not relevant in the cell line that has been used (Figure 4A and data not shown). Strikingly, we found that activity of the luciferase reporter fused to the CDC2L6 (CDK11) 3′UTR is increased when the endogenous ACA45 small RNA is inhibited. The CDC2L6 gene product is a component of the mediator complex and, therefore, important for transcription (Conaway et al., 2005). For further validation of ACA45 sRNA effects on CDC2L6 expression, we mutated all predicted ACA45 sRNA-binding sites in the CDC2L6 3′UTR (Figures 4B and S1). Indeed, a luciferase reporter containing the mutated CDC2L6 3′UTR was not upregulated when endogenous ACA45 sRNA was inhibited (Figure 4B), indicating that ACA45 sRNA seed sequence matches are important for CDC2L6 expression.

In summary, our data demonstrate that ACA45 is processed to a small RNA that can function like a miRNA on the endogenous target CDC2L6, identifying the ACA45 sRNA as a potential transcriptional regulator in human cells.

Cellular snoRNAs with miRNA Processing Signatures

The intriguing finding that ACA45 can function like a miRNA prompted us to analyze processing of other snoRNAs. We generated small RNA libraries from human Ago1–4 complexes and mapped the sequence reads to snoRNAs (the detailed composition of the Ago1–4 libraries are currently analyzed and will be published elsewhere). We find reads originating from stem-loop structures within the snoRNAs ACA47, ACA36b, U92, HBI-100, ACA56, ACA3, and ACA50 (Figure 6). Both arms of the individual stems are present in the libraries, and the sequence with the lower abundance is indicated as ''star'' sequences in Table S1 (see also Tables S2 and S3 for individual snoRNA-derived sequence reads and read lengths). Our data obtained from larger sequencing data sets suggest that processing of snoRNAs to functional small RNAs is not unique to ACA45 and can be observed for other snoRNAs as well.

DISCUSSION

snoRNAs form a highly abundant class of noncoding RNAs in many different organisms. snoRNAs localize to the nucleolus and guide specific modifications of rRNAs or snRNAs (Matera et al., 2007). Moreover, snoRNAs have also been implicated in alternative splicing events (Kishore and Stamm, 2006). Here, we show that the snoRNA ACA45 is processed to a small RNA that can function like a miRNA. ACA45 processing is independent of the Drosha-containing microprocessor complex but requires Dicer. At least in vitro, Dicer can process the full-length ACA45, although it does not structurally represent a classical Dicer substrate. In northern blots, however, the strongest signal originates from the full-length ACA45, and only a minor portion is processed to a small miRNA-like RNA. This observation is consistent with the finding that ACA45 exists as a functional snoRNA that forms snoRNPs with the protein factor GAR-1 (Matera et al., 2007). Therefore, we propose a model in which ACA45 is transcribed and functions in the nucleolus of human cells (Figure 5). However, a minor portion is transported to the cytoplasm by a so far unknown export receptor. In the cytoplasm, Dicer immediately processes the full-length snoRNA to a miRNA-like small RNA that functions in gene silencing. This hypothesis is supported by our finding that recombinant Dicer, as well as Dicer-containing Ago protein complexes, are capable of generating ACA45 small RNAs in vitro. However, it cannot be excluded that other nucleases contribute to ACA45 processing in the cytoplasm. Alternatively, ACA45 is cleaved in the nucleus already, and one half is recognized as miRNA precursor by the miRNA pathway. However, such a scenario might be unlikely because a nuclear

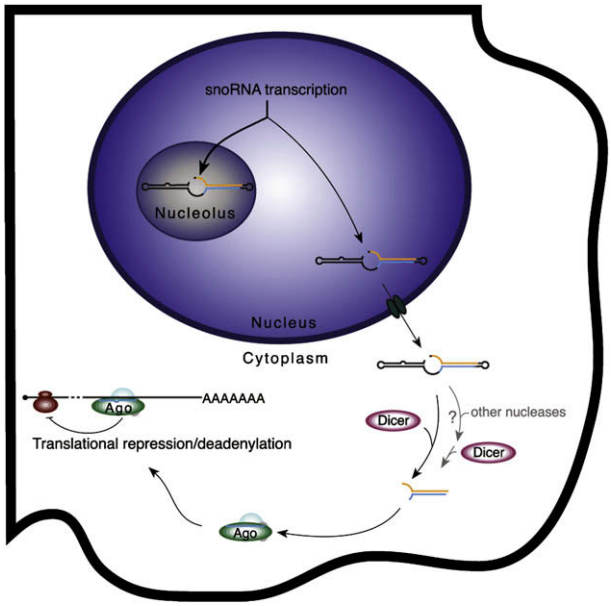


Figure 5. Model for ACA45 Processing and ACA45 sRNA Function in Human Cells

ACA45 snoRNA is transcribed in the nucleus, and the majority localizes to the nucleolus where it fulfills its specific functions by modifying other noncoding RNAs. However, a minor portion of ACA45 is exported to the cytoplasm, where Dicer, probably with the help of other nucleases, processes it to small RNAs that are specifically loaded into Ago protein-containing complexes. The ACA45-derived small RNA guides Ago protein complexes to partially complementary binding sites in the 3′UTR of target genes and represses its expression. AAAA, poly(A) tail.

cleavage activity might cleave the majority of the ACA45 pool, which is needed for classical snoRNA functions. Moreover, only one half of ACA45 would be exported by this model, although the other half folds like a typical miRNA precursor as well. Alternatively, a potential nuclear snoRNA cleavage activity could be physically separated from the snoRNAs as well. Further experiments aiming at the identification of specific snoRNA export pathways will help to elucidate the biogenesis pathways of small RNAs derived form snoRNAs.

Many small RNA cloning and sequencing projects have been carried out, but small RNAs derived from snoRNAs or other noncoding RNAs have not been reported. Here, we have immunoprecipitated endogenous Ago complexes, and it is very likely that small RNAs that associate with Ago proteins are functional RNA molecules rather than just degradation products. Most published cloning approaches size fractioned total RNA before cloning and, therefore, all unspecific degradation products are present in the libraries and it is difficult to find classes of functional small RNAs. Therefore, we suggest that cloning projects aiming at the identification of new classes of Ago-associated small noncoding RNAs of about 18–35 nt in length should be carried out from anti-Ago immunoprecipitations.

Using cloning and sequencing approaches, a variety of different snoRNA genes have been identified in the past (Bachellerie et al., 2002). However, many of these snoRNA candidates have not been characterized in detail, and it is unknown whether or not these candidates represent functional snoRNAs. Therefore, it is tempting to speculate that more snoRNAs are specifically processed to functional small RNAs. Indeed, by analyzing larger data sets, we find several small RNAs with miRNA-like processing signatures that originate from snoRNAs (Figure 6). These candidate sRNAs are derived from a subset of snoRNAs comprised of H/ACA snoRNAs and small Cajal body RNAs (scaRNAs), whose secondary structure is characterized by two hairpins linked by a hinge similar to ACA45 (Figure 2A). These findings support our hypothesis that a considerable number of snoRNAs are natural precursors for functional small RNAs. Moreover, we add another so far unrecognized function in post-transcriptional gene silencing to the list of snoRNA functions. A detailed functional characterization of all mammalian snoRNAs will help to elucidate the impact of snoRNA processing in RNA-guided gene silencing.

EXPERIMENTAL PROCEDURES

Ago Complex Purification

HEK293 cells were lysed in buffer containing 20 mM Tris HCl (pH 7.5), 150 mM NaCl, 0.25% NP-40, and 1.5 mM $MgCl_2$ and centrifuged at 10,000 × g for 10 min at 4°C.

For immunoprecipitation of endogenous Ago complexes, 100 μl protein G Sepharose (GE Healthcare) was washed with phosphate-buffered saline (PBS) and incubated with 10 ml anti-Ago1-4B8, anti-Ago2-11A9, anti-FLAG-3H3, or anti-GST at 4°C with gentle agitation overnight. After washes with PBS, beads were incubated with HEK293 cell lysate of 6 × 15 cm plates for 3 hr. Anti-Ago1-coated beads were extensively washed with 300 mM NaCl, 2.5 mM $MgCl_2$, 0.5% NP40, and 20 mM Tris-HCl (pH 7.5) followed by a wash with PBS. Anti-Ago2-coated beads were washed five times using RIPA buffer (50 mM Tris-HCl, 500 mM NaCl, 1% Nonidet P-40, 0.5% sodium deoxycholate, 0.1% SDS). RNA was isolated with 40 μg Proteinase K in 200 μl Proteinase K buffer (300 mM NaCl, 25 mM EDTA, 2% SDS, 200 mM Tris HCl [pH 7.5]) followed by Phenol/Chloroform extraction and Ethanol precipitation.

For immunoprecipitation of FLAG/HA-tagged Ago complexes, cell lysate from two 15 cm dishes were incubated with 20 μl FLAG M2 agarose beads (Sigma) for 2 hr at 4°C with rotation. Beads were extensively washed, and coimmunoprecipitated RNA was extracted as described above.

Small RNA Cloning

Small RNA cloning was carried out by Vertis Biotechnology (Weihenstephan, Germany) and has been described earlier (Tarasov et al., 2007). Without any size fractionation, extracted Ago-associated RNAs were poly(A)-tailed using poly(A) polymerase, and an adaptor was ligated to the 5′ phosphate of the miRNAs: (5′ end adaptor [43 nucleotides]: 5′-GCCTCCCTCGCGCCATCAG CTNNNNGACCTTGGCTGTCACTCA-3′). NNNN represents a "barcode" sequence. Next, first-strand cDNA synthesis was performed using an oligo(dT)-linker primer and M-MLV-RNase H reverse transcriptase (3′ end oligo [dT] linker primer [61 bases]: 5′-GCCTTGCCAGCCCGCTCAGACGAGACATCGCCCCG $C[T]_{25}$-3′). The resulting cDNAs were PCR amplified in 22 cycles using the high-fidelity Phusion polymerase (Finnzymes). The 120–135 bp amplification products were confirmed by polyacrylamide gel electrophoresis (PAGE) analysis. Both cDNAs pools were mixed in equal amounts and subjected to gel fractionation. The 120–135 bp fraction was electroeluted from 6% PAA-gels. After isolation with Nucleospin Extract II (Macherey and Nagel), cDNA pools were dissolved in 5 mM Tris/HCl (pH 8.5) with a concentration of 10 ng/μl and used in single-molecule sequencing. Massively parallel sequencing was performed by 454 Life Sciences (Branford, USA) using the Genome Sequencer 20 system as well as MWG Biotech (Germany). The complete sequencing data is available at the Gene Expression Omnibus (GEO, Accession number: GSE13370.)

A

type	ID	tot. # reads	"mature" seq. [# of reads]	"star" seq. [# of reads]	3' over -hang
scaRNA	ACA45	3516	aagguagauagaacaggucuugu [3405]	agaccuacuuaucuaccaacagc [55]	Y
scaRNA	ACA47	136	auuugcaguaacaggugugagc [107]	aucauguaugauacugcaaacag [6]	Y
scaRNA	U92	9	uaacggacagauacggggcagaca [5]	acugccuuuugaugaccgggacg [4]	Y
H/ACA Box RNA	HBI-100	591	uauggaggucucugucuggcuu [544]	ucugaucguucccccuccaua* [3]	Y
H/ACA Box RNA	ACA36B**	269	acuggcuagggaaaaugauugg [207]	uauucauuuauccccagccuaca [20]	N
H/ACA Box RNA	ACA56	102	aguggugaguucuucuguccagc [95]	ugguggcuuuagacuugccaga [2]	N
H/ACA Box RNA	ACA3	98	aucgaggcuagagucacgcuugg [87]	agugugcuagaguccucgaag [1]	Y
H/ACA Box RNA	ACA50	11	aagcacugccuuugaaccugaugu [8]	acgggccaagcaacagugcuaga [3]	Y (5nt)

B

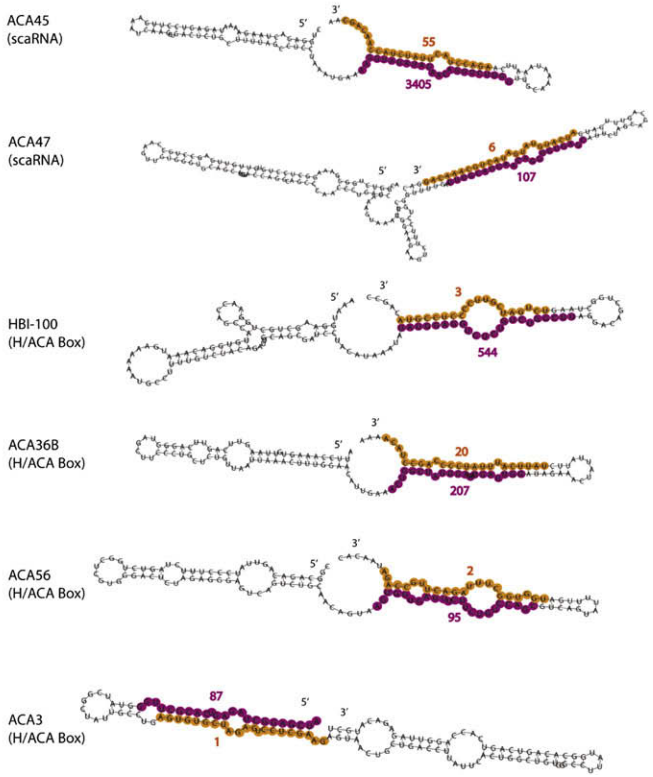


Figure 6. Several Human snoRNAs Show miRNA-Like Processing Signatures

(A) Small RNA reads originating from human snoRNAs that have been found in large sequencing data sets from Ago immunoprecipitates. The more abundant read is indicated as "mature," and the complementary strand is indicated as "star" read. All reads that have been found for individual snoRNAs are indicated as "total reads."

*The official genomic sequence is tctgatcgttcccctcc gta, but all of the reads mapping to this position have a mismatch, and all have "a" at position 18 and there is an annotated SNP at this position.

**The ACA36b sRNA candidate is identical to the annotated miRNA miR-664.

(B) Schematic representation of the secondary structure of full-length snoRNAs. Ago-associated reads are highlighted in purple and yellow.

RNA Cleavage Experiments

In vitro transcribed pri-27a substrate used in this study was described previously in Landthaler et al. (2004) and Meister et al. (2005). The template for pre-27a transcription was created by annealing the following primers: 5′-TTAATACGACTCACTATAGCTGAGGAGCAGGGCTTAGCTGCTTGTGAGCAGGGTCCACACCAAGTCGTGTTCACAGTGGCTAAGTTCCGCCCCCCAGC and 5′-GCTGGGGGGCGGAACTTAGCCACTGTGAACACGACTTGGTGTGGACCCTGCTCACAAGCAGCTAAGCCCTGCTCCTCAGCTATAGTGAGTCGTATTAA. ACA45 was cloned from genomic DNA using the primers 5′-ACGAGCTCCTGGAGACTAAGAAAATAGAGTCCTTGA and 5′-ACGGTACCTGCTGTTGGTAGATAAGTAGGTCTTGAA, digested with SacI and KpnI, and inserted into the SacI and KpnI restriction sites of the pBluescript. Plasmid was linearized using the KpnI restriction site and in vitro transcribed as described in Landthaler et al. (2004). The construction of human FLAG/HA-Ago1, FLAG/HA-Ago2, and FLAG/HA-Dicer was reported earlier (Meister et al., 2005). FLAG/HA-DGCR8 was purchased from Addgene.

Immunoprecipitations were performed as described above. For cleavage activity assays, 10 μl of Ago or Dicer complex-containing anti-FLAG beads were incubated in 20 μl PBS containing 5 mM ATP, 7.5 mM $MgCl_2$, 10 U/ml RNasin (Promega), and about 100 counts (~50 fmol) of internally labeled RNA for 1 hr at 37°C. The reaction was stopped by adding 200 μl proteinase K buffer (300 mM NaCl, 25 mM EDTA, 2% SDS, 200 mM Tris HCl [pH 7.5]) containing proteinase K (0.2 mg/ml). RNA was isolated with Phenol/Chloroform and analyzed by 8% or 12% denaturing RNA PAGE. Signals were detected by autoradiography.

Northern Blotting and Semiquantitative RT-PCR

Immunoprecipitated RNA and 30 μg total RNA isolated from HEK293 cells using Trifast (Peqlab) was separated by 12% denaturing RNA PAGE and transferred to a nylon membrane (GE Healthcare) by semidry electroblotting. Membranes were crosslinked by 1-ethyl-3-(3-dimethylaminopropyl) carbodiimide (EDC) chemical crosslink incubating for 1 hr at 50°C, prehybridized for 1 hr, and hybridized overnight at 50°C with probes complementary to snoRNA ACA45 or tRNA. The following probes have been used: 5′-AAGACCTGTTCTATCTACCT complementary to snoRNA ACA45 and 5′-CTGATGCTCTACCGACTGAGCTATCCGGGC complementary to lysine tRNA. After hybridization, membranes were washed twice 10 min with 5 × SSC and once 10 min with 1 × SSC. Radioactive signals were detected by exposure of BioMax MS film (Kodak) using an intensifying screen (GE Healthcare).

For semiquantitative RT-PCR, extracted RNA was modified by addition of poly(A) tail using poly(A) tailing kit from Ambion. Reverse transcription was performed using the cDNA synthesis kit (Fermentas) with the universal RT primer 5′ AACGAGACGACGACAGACTTTTTTTTTTTTTT (described in Hurteau et al. [2006]). DNA was amplified using Mesa Green qPCR MasterMix Plus (Eurogentec), a universal reverse primer identical to the 18 bp tag added during the RT step and the following specific primers: 5′-AAGGUAGAUAGAACAGGUCUUG for ACA45, 5′-TCCCTGAGACCCTAACTTGTGA for miR-125b, and 5′-ACACATCCAAATGAGGCG for 7SK. The PCR products were analyzed by 4% agarose gel electrophoresis.

Conserved Processing of ACA45

The human ACA45 sequence was obtained at the snoRNABase (http://www-snorna.biotoul.fr/). The ACA45 mouse, rat, and dog homologs were

identified by mapping the human sequence against each genome, retaining only unambiguous matches. Subsequently, a number of deep sequencing data sets were mapped to the ACA45 homologs. Each data set was mapped to the homolog of the species from which the data set originated, and only perfect matches were retained. The human data consisted of the data sets produced for this study using the 454 Life Sciences technology, as well as a data set produced by deep sequencing the small RNA fraction of HeLa cells using the Solexa/Illumina technology (GEO accession number GSE10829) (Friedländer et al., 2008). The mouse data sets were produced by deep sequencing small RNAs from mouse brain and kidney tissues using the 454 technology (unpublished data). The rat data set was produced by deep sequencing column-purified small RNAs from testes extracts using the 454 technology (GEO accession number GSE5026) (Lau et al., 2006). The dog data set was produced by sequencing small RNAs from dog lymphocytes using the Solexa technology (GEO accession number GSE10825) (Friedlander et al., 2008).

Computational Methods

A total of 64733 reads was obtained by deep sequencing the RNA that immunoprecipitated with Ago1 and Ago2. Of this, 20834 belonged to the Ago1 set and 43899 to the Ago2 set. Upon removal of adapters, the sequences shorter than 17 nt were discarded, resulting in 20448 and 42604 reads in Ago1 and Ago2 sets, respectively. These reads were mapped to human genome (hg 18, UCSC database [Karolchik et al., 2003]) using NCBI blastn (Altschul et al., 1990) with the minimum word length set to 7. The mapping with the best E value was associated with each read. The only mismatches allowed were the first nt at the 5′ end or the last three nt at the 3′ end of the read. In case a read mapped with the same E value to several locations, they were all taken into consideration. The genomic loci of best matches were annotated using the tables from UCSC database (Karolchik et al., 2003). A read was annotated as a DNA repeat (including LINE, SINE, LTR) only if the genomic locus it mapped to had no other annotation.

For purposes of identification of known and novel miRNAs, reads from the Ago1 and Ago2 libraries were combined and mapped to the human genome using NCBI megablast with the following options: −W 12 −p 100. Only perfect mappings (full length, 100% identity) were retained. These were used as input to miRDeep, an algorithm designed for the discovery of Dicer substrates such as miRNAs from deep sequencing data (Friedländer et al., 2008). The algorithm intersects the mappings with local genomic sequence to identify potential Dicer hairpin substrates. These are then scored according to the distribution of positions and frequencies of the reads mapped to the individual hairpin, using Bayesian statistics. The energetics and stability of the hairpins and the cross-species conservation of the seed sequence also contribute to the score. Human snoRNA sequences were downloaded from snoRNABase (Lestrade and Weber, 2006).

To map the total of 17362367 sequence reads obtained by sequencing Ago1–4 IP using Solexa technology to the genome, we used the locally developed suffix array-based tool (to be published elsewhere). Candidate snoRNAs with miRNA-like processing were selected (Table S1) if the combined Ago1–4 data set contained reads mapping to both strands of a hairpin and if these reads represented more than 85% of all reads mapping to a given snoRNA.

SUPPLEMENTAL DATA

The Supplemental Data include Supplemental Experimental Procedures, one figure, and three tables and can be found with this article at http://www.molecule.org/supplemental/S1097-2765(08)00733-8.

ACKNOWLEDGMENTS

We are grateful to Sabine Rottmüller and Bernd Haas for technical support, Vertis Biotechnologie AG (Weihenstephan, Germany) for small RNA cloning, and MWG biotech (Munich, Germany) for deep sequencing. We thank Sihem Cheloufi and Greg Hannon for providing RNA from Dicer$^{-/-}$ as well as Dicer$^{-/+}$ cells and Witold Filipowicz for the anti-GAR-1 antibody. Our research is supported by the Max Planck Society. This work was supported, in part, by EU grant LSHG-CT-2006-037900 (SIROCCO; G.M.) and the Deutsche Forschungsgemeinschaft (Me 2064/2-1 to G.M.).

Received: August 18, 2008
Revised: October 16, 2008
Accepted: October 27, 2008
Published: November 20, 2008

REFERENCES

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

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

Azuma-Mukai, A., Oguri, H., Mituyama, T., Qian, Z.R., Asai, K., Siomi, H., and Siomi, M.C. (2008). Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. Proc. Natl. Acad. Sci. USA *105*, 7964–7969.

Bachellerie, J.P., Cavaille, J., and Huttenhofer, A. (2002). The expanding snoRNA world. Biochimie *84*, 775–790.

Beitzinger, M., Peters, L., Zhu, J.Y., Kremmer, E., and Meister, G. (2007). Identification of human microRNA targets from isolated argonaute protein complexes. RNA Biol. *4*, 76–84.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature *409*, 363–366.

Bohnsack, M.T., Czaplinski, K., and Gorlich, D. (2004). Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA *10*, 185–191.

Borchert, G.M., Lanier, W., and Davidson, B.L. (2006). RNA polymerase III transcribes human microRNAs. Nat. Struct. Mol. Biol. *13*, 1097–1101.

Chen, K., and Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. Nat. Rev. Genet. *8*, 93–103.

Conaway, R.C., Sato, S., Tomomori-Sato, C., Yao, T., and Conaway, J.W. (2005). The mammalian Mediator complex and its role in transcriptional regulation. Trends Biochem. Sci. *30*, 250–255.

Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F., and Hannon, G.J. (2004). Processing of primary microRNAs by the Microprocessor complex. Nature *432*, 231–235.

Filipowicz, W., Jaskiewicz, L., Kolb, F.A., and Pillai, R.S. (2005). Post-transcriptional gene silencing by siRNAs and miRNAs. Curr. Opin. Struct. Biol. *15*, 331–341.

Friedländer, M.R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using miRDeep. Nat. Biotechnol. *26*, 407–415.

Gregory, R.I., Yan, K.P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. Nature *432*, 235–240.

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

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

Hurteau, G.J., Spivack, S.D., and Brock, G.J. (2006). Potential mRNA degradation targets of hsa-miR-200c, identified using informatics and qRT-PCR. Cell Cycle *5*, 1951–1956.

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

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. (2003). The UCSC Genome Browser Database. Nucleic Acids Res. *31*, 51–54.

Kishore, S., and Stamm, S. (2006). The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science *311*, 230–232.

Kiss, A.M., Jady, B.E., Bertrand, E., and Kiss, T. (2004). Human box H/ACA pseudouridylation guide RNA machinery. Mol. Cell. Biol. *24*, 5797–5807.

Krek, A., Grun, D., Poy, M.N., Wolf, R., Rosenberg, L., Epstein, E.J., MacMenamin, P., da Piedade, I., Gunsalus, K.C., Stoffel, M., and Rajewsky, N. (2005). Combinatorial microRNA target predictions. Nat. Genet. *37*, 495–500.

Landthaler, M., Yalcin, A., and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and its D. melanogaster homolog are required for miRNA biogenesis. Curr. Biol. *14*, 2162–2167.

Lau, N.C., Seto, A.G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D.P., and Kingston, R.E. (2006). Characterization of the piRNA complex from rat testes. Science *313*, 363–367.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V.N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415–419.

Lee, Y., Kim, M., Han, J., Yeom, K.H., Lee, S., Baek, S.H., and Kim, V.N. (2004). MicroRNA genes are transcribed by RNA polymerase II. EMBO J. *23*, 4051–4060.

Lestrade, L., and Weber, M.J. (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res. *34*, D158–D162.

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

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

Lund, E., Guttinger, S., Calado, A., Dahlberg, J.E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95–98.

Ma, J.B., Yuan, Y.R., Meister, G., Pei, Y., Tuschl, T., and Patel, D.J. (2005). Structural basis for 5′-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature *434*, 666–670.

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

Matera, A.G., Terns, R.M., and Terns, M.P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat. Rev. Mol. Cell Biol. *8*, 209–220.

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

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

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

Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev. *16*, 720–728.

Murchison, E.P., Partridge, J.F., Tam, O.H., Cheloufi, S., and Hannon, G.J. (2005). Characterization of Dicer-deficient murine embryonic stem cells. Proc. Natl. Acad. Sci. USA *102*, 12135–12140.

Parker, J.S., and Barford, D. (2006). Argonaute: a scaffold for the function of short regulatory RNAs. Trends Biochem. Sci. *31*, 622–630.

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

Patel, D.J., Ma, J.B., Yuan, Y.R., Ye, K., Pei, Y., Kuryavyi, V., Malinina, L., Meister, G., and Tuschl, T. (2006). Structural biology of RNA silencing and its functional implications. Cold Spring Harb. Symp. Quant. Biol. *71*, 81–93.

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

Pillai, R.S., Bhattacharyya, S.N., and Filipowicz, W. (2007). Repression of protein synthesis by miRNAs: how many mechanisms? Trends Cell Biol. *17*, 118–126.

Rudel, S., Flatley, A., Weinmann, L., Kremmer, E., and Meister, G. (2008). A multifunctional human Argonaute2-specific monoclonal antibody. RNA *14*, 1244–1253.

Seto, A.G., Kingston, R.E., and Lau, N.C. (2007). The coming of age for Piwi proteins. Mol. Cell *26*, 603–609.

Tarasov, V., Jung, P., Verdoodt, B., Lodygin, D., Epanchintsev, A., Menssen, A., Meister, G., and Hermeking, H. (2007). Differential regulation of microRNAs by p53 revealed by massively parallel sequencing: miR-34a is a p53 target that induces apoptosis and G1-arrest. Cell Cycle *6*, 1586–1593.

Tolia, N.H., and Joshua-Tor, L. (2007). Slicer and the argonautes. Nat. Chem. Biol. *3*, 36–43.

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

Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression

Marlon Stoeckius[1,4], Jonas Maaskola[1,4], Teresa Colombo[1,3], Hans-Peter Rahn[1], Marc R Friedländer[1], Na Li[1], Wei Chen[1], Fabio Piano[2] & Nikolaus Rajewsky[1]

***Caenorhabditis elegans* is one of the most prominent model systems for embryogenesis, but collecting many precisely staged embryos has been impractical. Thus, early *C. elegans* embryogenesis has not been amenable to most high-throughput genomics or biochemistry assays. To overcome this problem, we devised a method to collect staged *C. elegans* embryos by fluorescence-activated cell sorting (eFACS). In a proof-of-principle experiment, we found that a single eFACS run routinely yielded tens of thousands of almost perfectly staged 1-cell stage embryos. As the earliest embryonic events are driven by posttranscriptional regulation, we combined eFACS with second-generation sequencing to profile the embryonic expression of small, noncoding RNAs. We discovered complex and orchestrated changes in the expression between and within almost all classes of small RNAs, including microRNAs and 26G-RNAs, during embryogenesis.**

The nematode *Caenorhabditis elegans* is one of the best-explored model organisms for developmental biology. The mechanistic basis of embryogenesis in *C. elegans* has been dissected by describing the entire cell lineage[1] and by performing many molecular and genetic analyses. Various key proteins involved in early cell division as well as hundreds of essential genes required for early embryogenesis and their knockdown phenotypes have been described[2–8]. However, a true understanding of embryogenesis will require the knowledge of stage-specific gene expression. Modern high-throughput technologies such as deep sequencing, proteomics and their many applications can be used, for example, to identify and quantify the transcriptome, protein amounts and protein-protein interactions on a genome-wide scale. Prerequisite to the study of embryogenesis progression with many of these methods are large amounts of precisely staged embryos to yield enough RNA or other material. However, this is currently not possible. Isolated embryos are mixtures of embryos at developmental stages ranging from the early 1-cell zygote to the almost hatching worm larvae with approximately 600 cells. To date, staged embryos are usually obtained by manual sorting using a mouth pipette, making it impractical to apply large-scale techniques that require tens of thousands of embryos. Alternatively, one can obtain many semi-synchronized embryos by blocking their development with fluorodeoxyuridine[9], or one can isolate young embryos from hermaphrodites that have just begun to produce mature oocytes[10]. Although these methods can yield reasonable quantities of young embryos, the collected embryos are not synchronous, and these approaches cannot be used to investigate specific developmental stages.

Here we describe a method to collect many precisely staged embryos by fluorescence-activated cell sorting (eFACS). As *C. elegans* embryos have the same size throughout development, eFACS can in principle be applied to any embryonic stage in which a specific fluorescent marker protein can be stably expressed. Thus, eFACS allows the resolution of embryonic stages with sufficient yield of embryos for high-throughput analyses that require large amounts of starting material.

In *C. elegans* embryos, some zygote-specific transcription is initiated at the 4-cell stage, although pharmacological and genetic experiments have suggested that zygotic genes are not required until later in embryogenesis[11,12]. Maternal components seem sufficient to direct the embryo through the initial cleavage rounds up to approximately the onset of gastrulation. Interference with key enzymes involved in the RNA interference (RNAi) pathway lead to numerous defects including embryonic lethality, suggesting functional roles for noncoding RNAs in embryogenesis[13–15]. It is unknown which of the previously described small RNA populations in *C. elegans*[16–19] such as microRNAs (miRNAs), endogenous small interfering RNAs (siRNAs), 21U-RNAs (thought to be germline-specific and characterized by a length of 21 nucleotides (nt), a strong bias for 5′ uracils and their interaction with PIWI proteins) and the virtually uncharacterized class of 26G-RNAs (26 nt length and strong bias for a 5′ guanine) are present in the early embryo, and it is unclear how the complexity and composition of the small RNA transcriptome changes during the very first cell cycles[16–19]. We thus set out to use eFACS in combination with deep sequencing to profile small RNA expression during early embryogenesis.

[1]Max Delbrück Center for Molecular Medicine, Berlin, Germany. [2]New York University, Department of Biology and Center for Genomics and Systems Biology, New York, New York, USA. [3]Present address: Dipartimento di Biotecnologie Cellulari ed Ematologia, Sezione di Genetica Molecolare, Università La Sapienza, Rome, Italy. [4]These authors contributed equally to this work. Correspondence should be addressed to F.P. (fp1@nyu.edu) or N.R. (rajewsky@mdc-berlin.de).

RECEIVED 8 JUNE; ACCEPTED 10 AUGUST; PUBLISHED ONLINE 6 SEPTEMBER 2009; DOI:10.1038/NMETH.1370

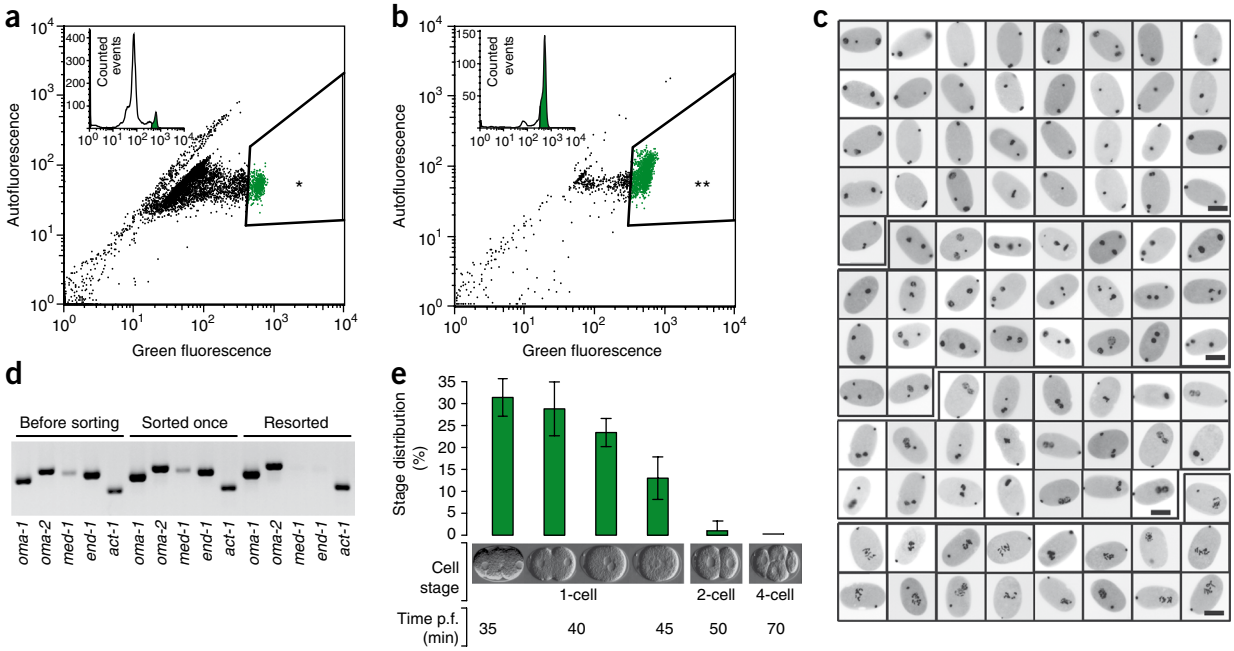


Figure 1 | eFACS yielded tens of thousands of staged 1-cell stage embryos with at least 98% purity. (**a**) Scatter plot of green fluorescence signal versus autofluorescence signal obtained by passing a mixed-stage population of embryos expressing an OMA-1–GFP fusion protein through a flow cytometer; 10,000 embryos are shown. Sorting 1-cell stage embryos with a high GFP signal (green; 3–7% of the initial population, *) yielded a sample enriched in 1-cell stage embryos (~70%). (**b**) Resorting this high GFP-positive population (green; 70% of the once-sorted population; **) yielded a virtually pure (>98%) 1-cell stage embryo sample; 2,000 embryos are shown. (**c**) Microscopy analysis of a randomly picked eFACS sample of 96 resorted 1-cell stage embryos. DAPI-stained pronuclei appear black. Images are grouped by embryo progression through the first cell cycle. Scale bars, 25 µm. (**d**) RT-PCR analysis of once-sorted embryos, resorted embryos and a mixed-stage embryo population before sorting. In resorted embryos, the 1-cell stage embryo–specific marker genes *oma-1* and *oma-2* but not early zygotic (4–8-cell stage) genes *med-1* and *end-1* were amplified. In once-sorted embryos expression of *med-1* and *end-1* was still detectable. *Act-1* was used as RT-PCR control. (**e**) Summary statistics of microscopy analyses of resorted embryos and timeline of early embryogenesis after fertilization (p.f.). The majority of embryos were in the pseudocleavage and chromosome condensation phase (30%) and in the pronuclear migration phase (30%) of the first cell cycle. Error bars, s.d. (n = 5); 100 embryos counted per replicate.

RESULTS

eFACS yields large samples of 1-cell stage embryos

The strain we used for eFACS experiments expresses an oocyte maturation factor 1 fused to GFP (OMA-1–GFP) under control of the *oma-1* promoter[20]. The OMA-1–GFP fluorescence is detected in developing oocytes and the 1-cell stage embryo. The GFP signal rapidly decreases in the two-cell embryo and is too weak to be detected in the embryo after the 4-cell stage[20]. These characteristics make the strain useful for selecting 1-cell stage embryos by fluorescence.

We collected a mixed-stage embryo population from gravid hermaphrodites of the *OMA-1*::*GFP* strain by standard methods[21]. We analyzed these embryos by flow cytometry, and 3–7% of the embryos had high GFP signal (**Fig. 1a**). Selecting this population for sorting in a fluorescence-activated cell sorting (FACS) machine yielded a sample of ~70% 1-cell stage embryos contaminated with older embryos (**Fig. 1b**). We investigated whether we could sort twice (hereafter referred to as resorting) to obtain an even higher enrichment in 1-cell stage embryos. The first embryonic cleavages progress rapidly and allow a time window of only 40 min for sorting living embryos[1]. After this time, a mixed-stage embryo population is depleted of 1-cell stage embryos. Even with extensive cooling to delay cell division, we were unable to achieve additional enrichment. However, methanol fixation of embryos allowed additional enrichment of the desired population (**Fig. 1b**) and routinely yielded ~60,000 almost pure (>98%) 1-cell stage embryos (**Fig. 1c–e**). Most of those resorted 1-cell stage embryos were in the pronuclear migration and pseudocleavage part of the first cell cycle (**Fig. 1c**,**e**). Selecting a population with a lower GFP signal during eFACS and resorting this population yielded a mixture of 2–4-cell stage embryos with some contamination of 15% 1-cell, 5% 8-cell and <2% older stages (**Supplementary Fig. 1**).

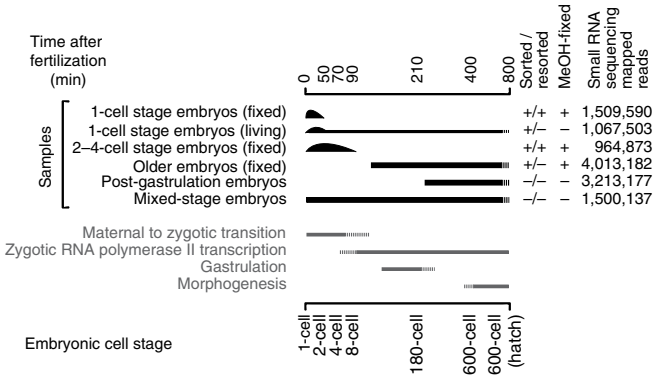


Figure 2 | Dissecting small RNA expression during embryogenesis. Summary of the six samples used for small RNA cloning and second-generation sequencing. Samples obtained by eFACS were sorted and resorted (+/+) or alternatively only sorted once (+/–) and fixed (+) or nonfixed live (–) sorted. Samples obtained by non-eFACS methods were not sorted (–/–) and not fixed (–). The total number of mapped small RNAs is given for embryos in the indicated stages.

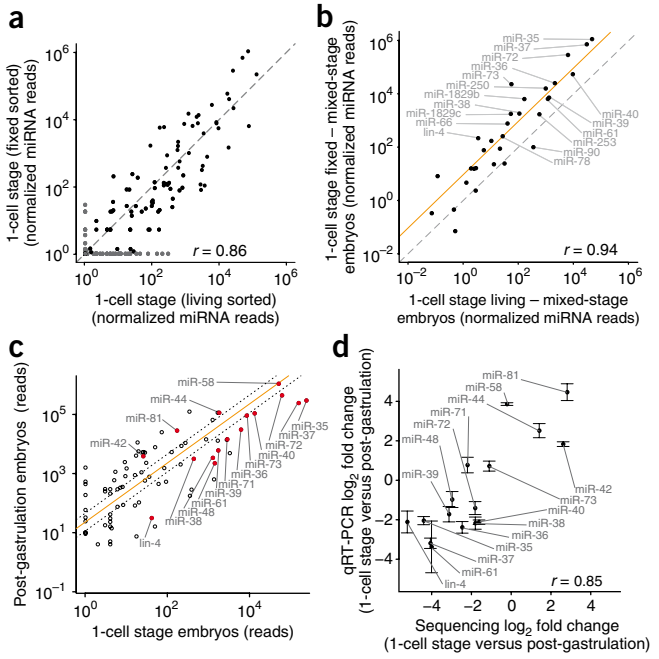


Figure 3 | Flow cytometry–based sorting of fixed 1-cell stage embryos revealed miRNA expression dynamics. (**a**) Scatter plot of miRNA expression quantified by normalized sequencing reads in the fixed resorted 1-cell stage embryos (>98% purity) versus the live-sorted sample (~70% 1-cell stage embryos and ~30% older embryos). miRNAs detected in only one sample are shown in gray. Dashed line, main diagonal. (**b**) miRNA expression in live- and fixed-sorted 1-cell stage embryos after *in silico* subtraction of miRNA expression in mixed-stage embryos. All miRNAs with higher expression in the 1-cell stage embryos compared to the mixed-stage embryos are labeled. Virtually all of these (linear regression, orange line) reside above the diagonal (dashed line) demonstrating the higher enrichment of 1-cell stage–specific miRNAs in the fixed-sorted sample. (**c**) Fold changes of miRNA expression between the 1-cell stage embryos and post-gastrulation embryos shown in a double-logarithmic scatter plot. Orange line, linear regression. Dotted lines, twofold change. (**d**) qRT-PCRs for selected miRNAs (marked in red in **c**) were performed to assay miRNA expression fold changes on living, hand-picked embryos. Expression fold changes obtained by sequencing were plotted versus expression fold changes obtained by qRT-PCRs. Coordinated expression fold changes were observed in miRNA clusters miR-35 and miRNA-42. Error bars, s.d. (n = 3).

Six samples covering different developmental stages

To study the composition and dynamics of small RNAs in early embryogenesis, we obtained six samples of embryos covering developmental stages from 1-cell stage to post-gastrulation embryos (**Fig. 2**). We generated 1-cell stage embryo samples from living and methanol-fixed embryos by eFACS with 70% and >98% purity, respectively (**Fig. 1**). We obtained a 2–4-cell stage embryo sample by eFACS of fixed embryos enriching for 2–4-cell stage embryos (**Supplementary Fig. 1**). We generated the two older embryo populations by (i) eFACS selecting the GFP-negative population of the *OMA-1*::*GFP* strain, which represents a mixed-stage embryo population depleted in early embryos, and by (ii) collecting post-gastrulation embryos by allowing isolated embryos to develop for 3 h at 20 °C. Finally, we also obtained an unsynchronized mixed-stage embryo population. We generated and deep sequenced small RNA libraries from all samples. Using our mapping pipeline (Online Methods), we mapped 52–83% of reads to the genome (**Fig. 2** and **Supplementary Table 1**).

Comparing fixed and living embryos obtained by eFACS

To first test whether methanol fixation altered miRNA expression, we compared expression profiles for 11 miRNAs between fixed and nonfixed embryos by quantitative reverse transcription–PCR (qRT-PCRs) (**Supplementary Fig. 2**). Relative expression of these miRNAs was unaffected by fixation. To compare the expression profile of fixed and living embryos after sorting, we examined sequencing-based estimates of miRNA expression between these samples. We expect some differences because we know that contrary to the fixed resorted eFACS sample, the living, once sorted, sample is contaminated by ~30% mixed-stage embryos.

We found that miRNA expression was overall highly correlated between these samples (**Fig. 3a**; Pearson correlation coefficient of log expression of 0.86), although we observed substantial scatter and some miRNAs that were absent in the fixed and resorted sample. We suspected that these miRNAs are expressed only in older embryos and were therefore not detected in the virtually pure fixed-resorted 1-cell stage embryo sample. To test this hypothesis, we first measured miRNA expression in an independently obtained, mixed-stage embryo sample (**Fig. 2**). We then subtracted miRNA expression values (Online Methods) of this mixed-stage embryo sample from miRNA expression values from both sorted samples (**Fig. 3b**). The expression values (estimated from sequencing data) of the small remaining set of miRNAs had (i) strongly reduced scatter between both sorted samples, (ii) correlated almost perfectly between these samples (Pearson correlation coefficient of log expression of 0.94), and (iii) were higher in the fixed sample. Thus, these miRNAs are likely 1-cell stage embryo–specific. Together, these data indicate that miRNA expression changes during embryonic development quantified by eFACS and deep sequencing can accurately reflect *in vivo* expression changes. However, because of sequencing biases, it is difficult to use sequencing-based estimates of miRNA expression to compare absolute *in vivo* expression between different miRNAs. We thus set out to analyze and validate miRNA expression using fold changes, in which sequencing biases largely cancel out.

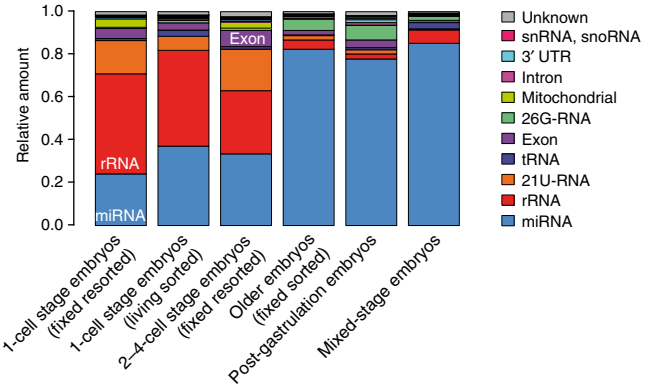


Figure 4 | Composition of different classes of small RNAs at different developmental stages obtained by deep sequencing. Small RNA categories were ordered by their overall abundance in all samples (miRNA > rRNA > 21U-RNA > tRNA > exons > 26G-RNA > mitochondrial > introns > 3′ UTR > sn- and snoRNA > unknown).

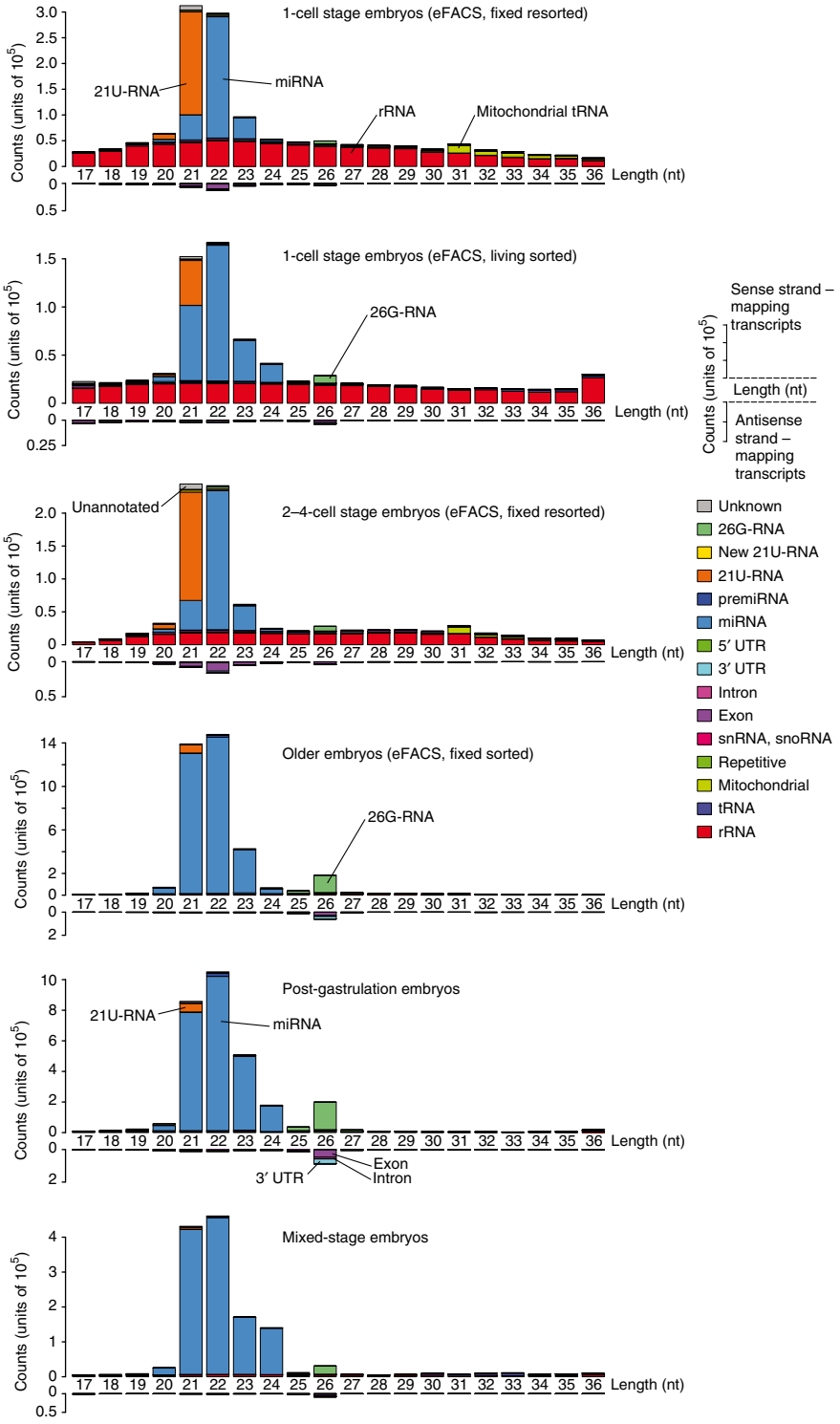


Figure 5 | Small RNA length distribution at different developmental stages. Stacked barplots show the number of reads of a given length mapping to the indicated feature categories.

Dynamics of miRNA expression in early development

eFACS revealed that ~60% of all known miRNAs are expressed in the 1-cell stage embryo (**Fig. 3a** and **Supplementary Table 2**). We selected 16 miRNAs with read counts covering three orders of magnitude for independent validation by qRT-PCR (**Fig. 3c**) on hand-picked, living 1-cell stage embryos and confirmed the expression of all of them (**Supplementary Table 3**). We then computed fold changes of miRNA expression from sequencing data between 1-cell stage embryos and our post-gastrulation sample according to a logistic model (Online Methods). We also directly assayed these miRNA expression fold changes by qRT-PCR for the 16 miRNAs on independently hand-picked, living embryos from corresponding developmental stages. miRNA expression fold changes determined by sequencing and qRT-PCR were well correlated (**Fig. 3d** and **Supplementary Fig. 3**; Pearson correlation coefficient (r) = 0.85). However, there were marked differences for some miRNAs (see Discussion). We next examined expression changes of all miRNAs between 1-cell stage, 2–4-cell stage and post-gastrulation embryo samples. The least amount of change was visible across the first cell divisions (1-cell stage to 2–4-cell stage embryos). However, miR-48 seemed to decrease greater than fivefold from 1-cell stage to 2–4-cell stage embryos (**Supplementary Fig. 4a**). We observed the strongest miRNA expression changes upon gastrulation, when several miRNAs were for the first time highly expressed (**Supplementary Fig. 4b**). Nevertheless, we also observed miRNAs that peaked in expression in the early embryo, including the miR-35 cluster (miR-35-41), the miR-61 cluster (miR-61 and miR-250) and miRNA-1829b/c. As noted above, we already observed these miRNAs to be enriched in the 1-cell stage embryo (**Fig. 3b**), and we conclude that these miRNAs are markers for very early embryogenesis.

Identification of new miRNAs and 21U-RNAs

To discover potentially new miRNAs, we mined our pooled datasets with miRDeep, an algorithm that identifies Dicer hairpin products such as miRNAs in deep-sequencing data[22]. miRDeep reported 19 new miRNAs (**Supplementary Table 4**); 16 were supported by detected star strands. Precursors of two new miRNAs fell exactly between adjacent coding exons, strongly suggesting that they are mirtrons. We observed expression from 7,506 of 15,341 known 21U-RNA loci (**Supplementary Tables 5,6**). Reads mapping to known 21U-RNA loci derived almost exclusively from the sense strand, had almost always a 5′ uracil, and their length distribution sharply peaked at 21 nt. We discovered 389 new 21U-RNAs (**Supplementary Table 7**).

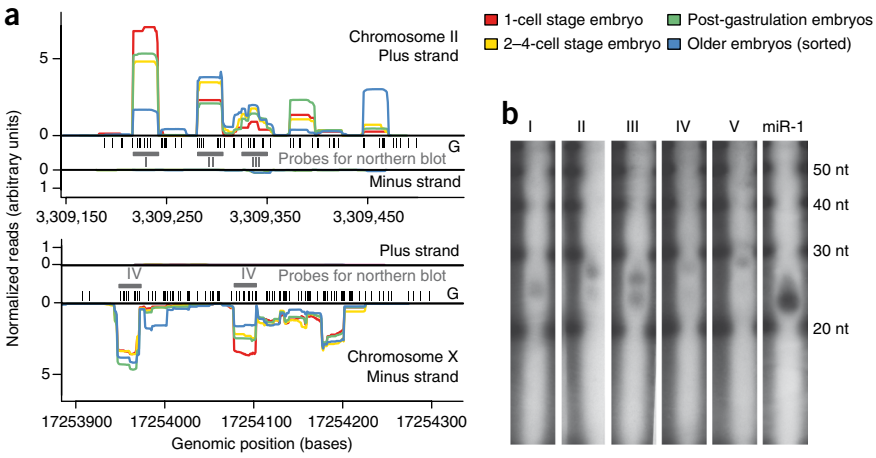


Figure 6 | The 26G-RNAs were expressed from intergenic clusters. (**a**) Analysis of two genomic clusters of 26G-RNAs on chromosome II and chromosome X. We observed most of the reads from one strand in our embryo samples. Black bars represent the positions of 5′ guanine nucleotides in the genome. Gray bars and roman numerals indicate the features for which northern blots were performed. Profiles are normalized to yield identical areas under the sense-strand curves. (**b**) Analysis of five 26G-RNAs by northern blots with total RNA from mixed-stage embryos. Roman numerals correspond to labels in **a**. The observed transcript lengths vary slightly between 24–28 nt. Some of the tested probes exhibit a double band in the northern blots (II, III, IV), and others (I and V) exhibit only a single band. A 21-nt probe for miR-1 was used as a positive control.

Their genomic distribution followed the published pattern[19,23] with additional dispersed genomic loci.

Differential expression across and within small RNA classes

We next compared the expression of all known classes of small RNAs during embryogenesis. However, we note that we most likely only observed small RNAs with a 5′ monophosphate owing to the cloning protocol. Overall, we observed strong, orchestrated changes in the composition of small RNAs between the sequenced samples (**Figs. 4**,**5**). Older embryos were dominated by miRNAs whereas in very early stages we observed additional small RNA classes. Those include mitochondrial tRNA as well as a sizable fraction of rRNA. The rRNA- and tRNA-derived fractions in all samples had a uniform length distribution and thus were likely to be degradation products. The 21U-RNAs were highly expressed in early embryos but difficult to detect in older embryos. We also observed differential expression of endo-siRNAs and 26G-RNAs. The relative abundance of small RNAs in mixed-stage embryo samples convoluted specific changes in small RNA expression during embryogenesis (**Figs. 4**,**5**).

Endogenous siRNAs are observed in the 1-cell stage embryo

The length distribution of reads mapping sense or antisense to exons or introns of mRNA transcripts varied distinctly (**Fig. 5**). Sense reads were distributed uniformly, suggesting that they originated from degraded mRNAs. Antisense reads mapping to exons were dominated by 22-nt and 26-nt reads with a strong bias for a 5′ uracil or guanine, respectively (consistent with previous reports[16,19]). We will refer to the corresponding small RNAs as endogenous siRNAs (endo-siRNAs). Most 1-cell stage embryo endo-siRNAs mapped to mitochondrial enzymes. The majority of these mRNAs are known to be upregulated in RNAi pathway defects (*rrf-1*, *eri-1*, *rde-3* and *dcr-1* mutants), which suggests that they are under control of small RNAs (**Supplementary Table 8**). We also consistently observed possible degradation products of mitochondrial tRNAs in the early embryo but not in other samples (**Fig. 5**). Notably, we found more ~22-nt endo-siRNA in the 1-cell stage and 2–4-cell stage embryos, whereas ~26-nt endo-siRNAs dominated in the older samples. Additionally, we observed in older embryos a twofold enrichment of antisense reads mapping to 3′ untranslated regions (UTRs) (27–32%) when compared to 1-cell or 2–4-cell stages (15%).

Genomic organization and expression of 26G-RNAs

After removing known RNA classes, we studied the set of remaining reads. The length distribution of these RNAs peaked at 26 nt and were most highly expressed in the older embryonic stages. These 26-mers did not map to any annotated loci and had a strong 5′ guanine bias (75.7%). Hereafter, we refer to 26-nt reads with a 5′ guanine as 26G-RNAs[24]. Although these 26G-RNAs were present only in low numbers in early embryos, we observed high 26G-RNA expression in older embryos. Computational analyses revealed that 26G-RNAs mapped to several clusters in intergenic regions on different chromosomes (**Fig. 6a**). We validated five (out of five tested) 26G-RNAs from two clusters (**Fig. 6b**).

DISCUSSION

In principle, eFACS can be used to extract large samples of embryos enriched in any desired embryonic stage. Thus, eFACS opens the door to many modern high-throughput technologies to assay embryonic stage–specific gene expression. Several of such investigations are already ongoing. A limitation of eFACS is that it depends on the availability of a good fluorescent marker gene for the desired embryonic stage. State-of-the-art flow cytometry analysis allows the simultaneous usage of up to eight fluorescence channels. Thus, strains expressing different fluorescent fusion proteins with temporally overlapping changes in gene expression could be combined, and thus it should be possible to use eFACS in situations in which a single optimal marker gene is not available. Moreover, protein stability could be tuned by engineering degradation at a specific time and in a specific cell type[25].

We sorted live embryos to obtain staged samples at a purity of ~70%. However, one technical constraint in eFACS is that the large size of the embryos forced us to sort at very low speeds of ~400 embryos per second. We cooled embryos (15 °C) to delay cell divisions but were still unable to resort living embryos. We also experimented with lower temperature settings (4–10 °C). However, these settings reduced viability (<60%) after sorting and still resulted in relatively low purity. It is entirely possible that more advanced flow cytometers will allow sorting at higher speeds comparable to that of standard cell sorting (>20,000 embryos per second). In this case, one could sort and even resort to obtain samples of the same size and purity as our fixed embryo eFACS runs. Fixation could be omitted and eFACS just with the *OMA-1*::*GFP* strain could be used to obtain thousands of staged living embryos that could be

allowed to develop synchronously to the embryonic stage of interest. Improvements to this approach might also be achieved by careful staging of worms[10] before eFACS.

We used methanol fixation before resorting embryos, and methanol fixation did not alter miRNA expression (**Supplementary Fig. 2**) or mRNA expression (data not shown). Nevertheless, we cannot rule out that methanol fixation or sorting does induce some artifacts when using eFACS for other purposes.

We observed some differences in miRNA expression fold changes determined by sequencing after eFACS or qRT-PCRs in hand-picked living embryos, including an outstanding discrepancy for miR-58. We believe that this discrepancy can be in part explained by saturation effects in the library preparation for sequencing because miR-58 is by far the most highly expressed miRNA. This problem and biases in sequencing in general may also be responsible for other discrepancies. Overall, we observed increased expression fold changes by qRT-PCR. An inherent problem when comparing sequencing and qRT-PCR data is that both methods require normalization. We normalized sequencing data under the assumption that net expression fold changes were close to zero whereas we normalized qRT-PCR results to an internal standard. Although both assumptions have their problems, different normalization procedures only shift the baseline of expression fold changes and do not influence the relative expression fold changes to each other and thus do not influence any conclusions presented in this study.

Previous large-scale studies of small RNA expression had used samples composed of mixed-stage embryos. These studies could not detect the orchestrated and dynamic changes between and within different classes of small RNAs that we observed when comparing the 1-cell stage embryos to later stages. First, the majority of miRNAs is already expressed in the 1-cell stage embryo, suggesting that they are maternally deposited. The reason remains to be determined. Second, we showed that miRNAs from the miR-35 cluster are likely early embryo–specific. Genetic knockouts and mutations for 95 miRNAs have been published[26]. Notably, the miR-35 cluster is the only known miRNA cluster with an embryonic lethal knockout phenotype. Third, we observed many small RNAs of uniform length mapping sense to rRNAs in 1-cell stage embryos (live-sorted or methanol-fixed), with decreased expression in 2–4-cell embryos, but virtually absent in samples from older stages. Thus, although we do not have independent validation, it seems unlikely that the observed rRNA expression is an experimental artifact. rRNAs, unlike mRNAs, are already transcribed in the 1-cell stage embryo[12]. One may speculate about a turnover of maternally and paternally provided rRNAs to zygotically transcribed rRNAs upon fertilization during very early embryogenesis. Finally, we found consistent evidence for a turnover of mitochondrial components in the 1-cell stage embryo. We observed degradation products of mitochondrial tRNAs in the early embryo as well as many siRNAs directed against mitochondrial enzymes. Thus, it is tempting to speculate about mechanisms that selectively degrade paternal mitochondria in early zygotes, as described in vertebrates[27].

Our data allowed us to study as yet virtually undescribed classes of small RNAs such as 26G-RNAs. Observations of small RNAs, in particular ~26-nt-long with a 5′ guanine bias have been reported earlier[16,19] and were recently dubbed 26G-RNAs[24]. We found that 26G-RNAs are dynamically expressed and that they cluster in several intergenic regions. Northern blot analysis suggested that they may be initially generated with heterogenous lengths or post-transcriptionally modified such that they appear as having different sizes on the northern blot. In addition to an increase in expression of 26G-RNAs in older embryos, we also observed increased expression of 26-nt endo-siRNAs mapping to the antisense strand of coding mRNAs. We did not computationally detect a 'ping pong' biogenesis mechanism[28,29] between 26G-RNAs and 26-nt endo-siRNAs.

Our eFACS data and analyses raise many more questions. However, altogether we are tempted to conclude that the complexity of small RNA expression dynamics in very early embryogenesis is comparable to the expression dynamics of protein-coding genes, and that the use of eFACS will contribute to a more complete understanding of gene regulatory networks during early animal development.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

Accession codes. Gene Expression Omnibus (GEO): GSE17153.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank R. Lin (University of Texas) for providing us with the TX189(P(oma-1)::oma-1::GFP) strain. All other strains used in this project were provided by the *Caenorhabditis* Genetic Center, which is funded by the US National Center for Research Resources. M.S. acknowledges part-time funding from the Berlin Institute for Medical Systems Biology, funded by Bundesministerium für Bildung und Forschung, and New York University PhD exchange program, and a travel grand from Boehringer Ingelheim Fonds. J.M. thanks the Deutsche forschungsgemeinschaft for a fellowship in the International Research Training Group Genomics and Systems Biology of Molecular Networks (GRK 1360). F.P. and N.R. acknowledge partial funding from US National Human Genome Research Institute (ModEncode U01 HG004276) and US National Institutes of Health (R01HD046236). We thank S. Lebedeva for help with sequencing runs.

AUTHOR CONTRIBUTIONS

F.P. and N.R. conceived, designed and supervised the study. M.S. designed and performed the experiments. J.M. designed and performed computational studies with the exception of predicting new miRNAs, which was done by M.R.F.; T.C. contributed to initial eFACS experiments; H.-P.R. helped with flow cytometer settings and runs; W.C. and N.L. contributed to library preparations and sequencing; M.S., J.M., F.P. and N.R. analyzed the data. M.S. and N.R. wrote the paper, and J.M. and F.P. edited it.

Published online at http://www.nature.com/naturemethods/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Sulston, J.E., Schierenberg, E., White, J.G. & Thomson, J.N. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**, 64–119 (1983).
2. Piano, F. *et al.* RNAi analysis of genes expressed in the ovary of *Caenorhabditis elegans*. *Curr. Biol.* **10**, 1619–1622 (2000).
3. Piano, F. *et al.* Gene clustering based on RNAi phenotypes of ovary-enriched genes in *C. elegans*. *Curr. Biol.* **12**, 1959–1964 (2002).
4. Kamath, R.S. *et al.* Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).
5. Gonczy, P. & Rose, L.S. Asymmetric cell division and axis formation in the embryo. *WormBook* **2005**, 1–20 (2005).
6. Sonnichsen, B. *et al.* Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* **434**, 462–469 (2005).
7. Fernandez, A.G. *et al.* New genes with roles in the *C. elegans* embryo revealed using RNAi of ovary-enriched ORFeome clones. *Genome Res.* **15**, 250–259 (2005).

8. Oegema, K. & Hyman, A.A. Cell division. *WormBook* **2006**, 1–40 (2006).
9. Stroeher, V.L. *et al.* DNA-protein interactions in the *Caenorhabditis elegans* embryo: oocyte and embryonic factors that bind to the promoter of the gut-specific *ges-1* gene. *Dev. Biol.* **163**, 367–380 (1994).
10. Schauer, I.E. & Wood, W.B. Early *C. elegans* embryos are transcriptionally active. *Development* **110**, 1303–1317 (1990).
11. Edgar, L.G., Wolf, N. & Wood, W.B. Early transcription in *Caenorhabditis elegans* embryos. *Development* **120**, 443–451 (1994).
12. Seydoux, G. & Dunn, M.A. Transcriptionally repressed germ cells lack a subpopulation of phosphorylated RNA polymerase II in early embryos of *Caenorhabditis elegans* and *Drosophila melanogaster*. *Development* **124**, 2191–2201 (1997).
13. Denli, A.M. *et al.* Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**, 231–235 (2004).
14. Grishok, A. *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**, 23–34 (2001).
15. Knight, S.W. & Bass, B.L. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in *Caenorhabditis elegans*. *Science* **293**, 2269–2271 (2001).
16. Ambros, V. *et al.* MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* **13**, 807–818 (2003).
17. Baugh, L.R. *et al.* Composition and dynamics of the *Caenorhabditis elegans* early embryonic transcriptome. *Development* **130**, 889–900 (2003).
18. Evans, T.C. & Hunter, C.P. Translational control of maternal RNAs. *WormBook* **2005**, 1–11 (2005).
19. Ruby, J.G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193–1207 (2006).
20. Lin, R. A gain-of-function mutation in oma-1, a *C. elegans* gene required for oocyte maturation, results in delayed degradation of maternal proteins and embryonic lethality. *Dev. Biol.* **258**, 226–239 (2003).
21. Brenner, S. The genetics of *Caenorhabditis elegans*. *Genetics* **77**, 71–94 (1974).
22. Friedlander, M.R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.* **26**, 407–415 (2008).
23. Batista, P.J. *et al.* PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in *C. elegans*. *Mol. Cell* **31**, 67–78 (2008).
24. Ghildiyal, M. & Zamore, P.D. Small silencing RNAs: an expanding universe. *Nat. Rev. Genet.* **10**, 94–108 (2009).
25. Nance, J., Munro, E.M. & Priess, J.R. *C. elegans* PAR-3 and PAR-6 are required for apicobasal asymmetries associated with cell adhesion and gastrulation. *Development* **130**, 5339–5350 (2003).
26. Miska, E.A. *et al.* Most *Caenorhabditis elegans* microRNAs are individually not essential for development or viability. *PLoS Genet.* **3**, e215 (2007).
27. Sutovsky, P. Ubiquitin-dependent proteolysis in mammalian spermatogenesis, fertilization, and sperm quality control: killing three birds with one stone. *Microsc. Res. Tech.* **61**, 88–102 (2003).
28. Aravin, A.A. *et al.* Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**, 744–747 (2007).
29. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **128**, 1089–1103 (2007).

ONLINE METHODS

Strains. We used wild-type *C. elegans* (N2) and the TX189(P(oma-1)::oma-1::gfp) strain for sorting early embryos. Strains were maintained using standard methods[21] on OP50-seeded NGM plates at 20 °C unless otherwise noted.

Liquid culture. Liquid culture of *C. elegans* was modified from previous protocols[30]. Worms were cultivated in S-Basal medium (100 mM NaCl, 6 mM K_2HPO_4, 44 mM KH_2PO_4 and 5 mg l^{-1} cholesterol) supplemented with 3 mM $MgCl_2$, 3 mM $CaCl_2$ and 10 mM K-citrate (pH 6) on a rotary shaker at 180 r.p.m. The liquid culture medium (S-medium) had a pH of ~ 6 and an osmolarity of around 370 mOsmol kg^{-1}.

Detailed eFACS protocol. A fluorescent protein fusion strain is needed for this procedure, which expresses the fluorophore at the desired cell stage of interest. The required fluorescence intensity for sorting is only limited by the lowest concentration necessary to distinguish between labeled and unlabeled populations. Differences below 10% in GFP signal level are still picked up by a sorter because of the photomultipliers, which increase the sensitivity of the flow cytometer by several orders of magnitude[31]. We still obtained successful sorts when picking gates such that we had only a ~50% difference in GFP signal levels.

The strain we used for sorting expressed an OMA-1–GFP fusion protein under the control of the *oma-1* promoter[20]. The GFP fluorescence was detected in the developing oocytes and the 1-cell stage embryo, decreased below 10% in the 2-cell embryo and was too weak to be detected in the embryo after the 4-cell stage[20]. These characteristics made the strain very useful for selection of 1-cell-stage embryos by fluorescence.

Millions of worms were needed for this procedure. Synchronized worms should optimally be grown in liquid culture to reduce the amount of space needed for growth. We routinely clean our worms on Ficoll 400 (1.077 g ml^{-1}) (PAA Laboratories) from bacteria and precipitate in the liquid culture. Embryos should be extracted from young adults to minimize the amount of older embryos in the population. Embryos have to be isolated under cooled conditions (4 °C M9 buffer and centrifuge) and as fast as possible. Embryos were extracted as described previously[21] with higher percentage sodium hypochloride bleach (12%) (Carl Roth) to ensure faster release of eggs. This bleaching protocol does not alter viability of embryos but reduces bleaching time and contamination with worm debris. Worms were monitored under the dissecting microscope during the bleaching procedure, and the procedure was stopped as soon as most worms were dissolved in the bleach. The eggs and debris were then washed (310*g* for 30 s) twice in cold M9 and filtered through a 40-µm nylon mesh (Cell strainer; Falcon) to clean embryos from worm debris. Eggs were then resuspended in cold PBS (pH 7.4) and pelleted (310*g* for 30 s). The supernatant was discarded and embryos were fixed by resuspending them in −20 °C methanol (80%). The tube was then placed on an overhead rotator at 4 °C for at least 1 h.

The fixed embryos were pelleted (310*g* for 1 min) and resuspended in cold cell-culture grade PBS. Embryos were cleaned from large embryo aggregates and worm debris by passing the embryos through a 40-µm nylon mesh. Embryos have to be kept on ice in the tube and aliquots are sorted stepwise. Embryos tend to aggregate. Shortly before sorting they should be vortexed vigorously and passed though a 40-µm nylon mesh into a flow cytometer tube.

Fixed embryos were sorted on a FACSVantage SE (Becton Dickinson Inc.) using the 100 µm nozzle with a pressure of 8 p.s.i. and 14,600 Hz frequency. GFP was excited with an argon-ion laser (488 nm) and detected using the FL1 parameter (emission filter: 530 ± 15 nm) in comparison to the FL2 parameter (emission filter 585 ± 21 nm). Debris in the sample was excluded from the sort by gating for intact embryos using the forward and side scatter. Nonfluorescent wild-type (N2) embryos were included as control. Data were analyzed using 10,000 events per sample for the first sort and 2,000 events per sample for the resort using CellQuest Software (BD Biosciences) (**Fig. 1a**,**b**).

To prevent embryos from aggregating in the flow cytometer tube during the run, we introduced a magnet stirrer on the bottom of the tube, which was triggered by a remote-controlled magnetic stirring device (Variomag; Thermo Electron). Embryo aggregates will clog the nozzle of the flow cytometer. The stirrer also prevents embryos from settling at the bottom of the tube, which ensures a constant distribution of embryos in the suspension. We had best results at a sorting speed between 400–500 events (embryos) per second; increasing sorting speed decreases efficiency. Embryos were sorted into a siliconized dish filled with PBS. Embryos should be sorted twice (hereafter referred to as 'resorted') to ensure a sufficiently high purity. In the resort, the embryos were sorted directly into a dish of Trizol LS (Invitrogen). Optionally embryos can be sorted into PBS and collected by centrifugation before the desired downstream application.

During the second flow cytometer run, we prepared several microscope slides by dropping at least 200 embryos onto each slide. Slides were covered with mounting medium containing DAPI nuclear stain (Vectashield; Vector Laboratories). This allowed examination of the purity of the resorted embryo sample by microscopy. Additionally the purity was determined by PCR-amplifying maternal 1-cell stage (*oma-1* and *oma-2*) and zygotic >4–8-cell stage (*end-1* and *med-1*) marker genes.

eFACS was used in this study to isolate 1-cell stage embryos (high GFP-positive population; **Fig. 1**), 2–4-cell stage embryos (intermediate GFP-positive population; **Supplementary Fig. 1**) and older embryos (GFP-negative population).

Isolation of post-gastrulation embryos. Gravid wild-type (N2) adults were treated with sodium hypochloride bleach to extract embryos as described previously[30]. In *C. elegans,* gastrulation is initiated at approximately the 25-cell stage. Embryos were allowed to develop in S-medium at 20 °C for 3 h. At this time all embryos have gastrulated, which we confirmed under a dissecting microscope. Hatched L1 larvae were removed by a second round of sodium hypochloride bleach treatment. We allowed 100 embryos to hatch on NGM plates overnight to determine viability.

RNA extraction. RNA was isolated from all samples by two rounds of freeze-thaw lyses of the embryos in Trizol LS reagent (Invitrogen). RNA was precipitated with Glycoblue (Ambion) overnight at −20 °C or for 30 min at −80 °C.

Deep sequencing ('Illumina'). Library preparation as well as cluster generation and deep sequencing were performed according to the 5′ ligation–dependent (5′ monophosphate–dependent)

doi:10.1038/nmeth.1370

manufacturer's protocol (Digital Gene Expression for small RNA; Illumina). Roughly 60,000 embryos (~10 µg total RNA) were used for small RNA library preparation. Small RNAs were size-selected between 18 and 40 nt according to the single-stranded DNA marker in the small RNA sequencing kit (Illumina). Small RNA libraries from the early embryo samples (1-cell stage and 2–4-cell stage) as well as the mixed embryos and older sorted embryos were sequenced on the Genome Analyzer 1 (Illumina), and the libraries generated from post-gastrulation embryos were sequenced on the Genome Analyzer 2 (Illumina).

Mapping. Mapping was performed using an in-house developed pipeline (J.M. *et al.*, unpublished data). This pipeline consists of an initial 3′ adaptor removal step, low-complexity read filtering, a mapping routine using a suffix-array based alignment program and a 3′ adaptor identification refinement phase.

Briefly, initial adaptor removal was performed by using dynamic programming to find in each read the suffix that best matched to a prefix of the 3′ adaptor. For this, all alignments of adaptor prefixes to suffixes of the read sequence were considered. In addition, occurrences of the full adaptor sequence anywhere in the read sequence were considered. Among these alignments, the best alignment was determined according to a simple one-parameter model $p(\text{alignment} \mid \Theta) = \Theta^n (1 - \Theta)^{n-k}$, where n is the length of the alignment, k is the edit distance of the alignment, and Θ is a parameter describing the error rate. A Θ value of 0.9 was heuristically chosen to reflect the relatively high error rate toward the end of Illumina reads.

The alignment program proceeds by determining all genomic matches to a read in edit distance k. For this application edit distance two was used. The alignment algorithm was implemented using a suffix array of the genome against which each read is sought, incrementally increasing the edit distance until matches are found.

In the 3′ adaptor identification refinement phase, the boundary between transcript and adaptor parts of each read was redetermined in light of the genomic context that the read was mapped to. This was done by computing a score $S(i) = f(i) + r(i)$ for every position i of the read. $f(i)$ is derived by aligning prefixes of the genomic context to prefixes of the read, from which $f(i)$ gives the edit distance of the best match of the read prefix of length i to the genomic context. $r(i)$, the second part of the score is determined from reverse alignments of the reversed read to reverse adaptor prefixes, that is, $r(i)$ was the edit distance of the best match between the read sequence positions $i + 1$ to n and a adaptor prefix, where n was the length of the read. The 3′ adaptor beginning position t was then determined so as to minimize $S(t)$. In case of ties, the minimum of the tied positions was used.

For the subsequent analyses we used weighted matches, that is, reads mapping to multiple loci have equal weight distributed across these loci (for example, a read represented by two transcripts and mapping equally well to three loci had a weight of two-thirds assigned to each of the three loci).

Normalization of miRNA reads in between samples. Expression fold changes of sequencing data were determined using linear models of the log expression, that is, logistic expression models, as follows. Assume we are given expression vectors $\mathbf{a},\mathbf{b} \in \mathbb{R}_{\geq 0}^{n}$, where n is the number of genes and $\mathbb{R}$ is the set of real numbers, with $\mathbf{a} = (a_i)_{i=1..n}$ and a_i the expression of gene i. If reference values $z_a, z_b \in \mathbb{R}$ are known, then the normalized expression values are $\mathbf{a}/z_a$ and $\mathbf{b}/z_b$, and the fold change $fc_i \in \mathbb{R}$ of gene i is given by

$$fc_i = \frac{\frac{b_i}{z_b}}{\frac{a_i}{z_a}} = \frac{b_i z_a}{a_i z_b} = \text{constant} \times \frac{b_i}{a_i}.$$

Here 'constant' denotes an arbitrary constant determined by the ratio of the two unknowns a_i and b_i. Thus, the log fold changes are

$$\log fc_i = \log b_i - \log a_i + \log z_a - \log z_b = \log b_i - \log a_i + \text{constant},$$

which is equivalent to

$$\log b_i = \log a_i + \log fc_i + \text{constant} \quad (1)$$

Typically, the expression of a RNA species that is known to be constant between the two samples is used for the reference values. However, for the present study no such constants were known. We resorted to fitting a linear model of the form response = predictor + residual + intercept to equation 1, in which $\log a_i$ is the predictor, $\log b_i$ the response, the intercept term determines the ratio of normalizers, and finally the log fold changes correspond to the residuals. In fitting, the slope of the linear model is fixed to unity, essentially only fitting the intercept term. From the fitted models the log fold changes are found as the prediction residuals. Owing to the slope of unity, it is possible to trivially accumulate the pairwise intercept terms of a sequence of expression samples for a joint normalization.

The proposed normalization method is equivalent to assuming that the mean log fold change is zero. The calculated fold changes of miRNA expression by this normalization method were validated by qPCRs and showed to be in good agreement (see below; **Fig. 4**).

Validation of observed miRNA expression patterns by TaqMan miRNA qPCR assays. The fold changes that were computed for 16 miRNAs from the deep sequencing data between 1-cell stage embryos and post-gastrulation embryos were validated by TaqMan miRNA qRT-PCR assays (Applied Biosystems) on hand-picked 1-cell stage embryos and older post-gastrulation embryos. Embryos were collected from cut gravid hermaphrodites by mouth pipette and washed thoroughly before lysing them in Trizol LS reagent. MicroRNA TaqMan PCR assays were performed following the recommendations of the manufacturer (Applied Biosystems). A TaqMan assay for the small RNA U18 and sn2343 was used as a normalization standard.

Classification, annotation and quantification of small RNA deep sequencing reads. Known miRNA coordinates were retrieved from miRBase release 12 (ref. 32). Other noncoding RNA, 3′ UTR, 5′ UTR, exon and intron coordinates were retrieved from WormBase, matching genome release WS190 (http://www.wormbase.org; WS190). The 21U-RNA coordinates were retrieved from the supplementary materials of previous studies[19,23]. Coordinates of RepeatMasker annotations[33] and simple repetitive sequences[34] were obtained from the UCSC genome browser[35]. Reads were

annotated by intersecting the mapped coordinates subsequently with the following sets of annotated feature coordinates and subtracting intersecting coordinates before proceeding with the next annotation set. The order in which annotation categories were used was: (i) miRNA, (ii) rRNA, (iii) tRNA, (iv) snRNA, (v) snoRNA, (vi) 21U-RNA, (vii) mRNA and (viii) repetitive sequences. This order roughly reflects the number of genomic bases represented by the different classes. For each annotated feature, all overlapping mapped reads were determined, and quantification was done by summing the overlapping weighted matches.

Annotating new 21U-RNAs. New 21U-RNAs were predicted using the motif scoring modules provided with ref. 19, which uses position-specific nucleotide frequency matrices of the large and small 21U-RNA upstream motifs, as well as a model for the distance between the two motifs to determine occurrences of 21U-RNA loci. The matrices are parameterized from ungapped alignments of reads deriving from manually selected portions of the genome that are rich in 21U-RNA. The motif scoring was applied to the set of mapped loci that remained after removing other known noncoding RNAs (including previously known 21U-RNA loci) in which the sequences scored consisted of the upstream 100 nt and the read itself. We used the same score cutoff of 15.5 to call loci as was used previously[19].

miRDeep analysis. To guide the excision of potential miRNA precursors from the genome, the above read mappings were used. First, we identified read mappings that (i) did not overlap with rRNAs, tRNAs or 21U-RNAs, (ii) were perfect mappings of edit distance zero, (iii) were from reads no shorter than 18 nt and no longer than 25 nt after removal of the adaptor and (iv) were from reads that did not have more than five perfect matches to the genome. Second, we identified genomic stacks of such reads ('stack' meaning two or more reads mapping to the same 5′ and 3′ positions). For each genomic locus, the highest read stack was identified and two potential precursor sequences excised, one spanning 20 nt upstream of the stack and 70 nt downstream of the stack and the other spanning 70 nt upstream and 20 nt downstream. This new excision algorithm will be part of the updated miRDeep package. Subsequently, all reads in our pooled datasets were remapped to these precursors using an edit distance of 1, and all suboptimal mappings were discarded (edit distance 1 matches for reads that have one or more perfect match). The read mappings to the potential precursors and the structures of the precursors were input to miRDeep as described previously[22]. For purposes of seed conservation, a limited set of known mature miRNAs from miRBase version 11 was used. These consisted of miRNAs from families that are present in invertebrates or that are conserved between mammalian and nonmammalian vertebrates. miRDeep initially reported 31 candidate miRNAs. These were manually curated to remove redundant sequences or highly palindromic precursors that were unlikely to represent genuine miRNAs.

In summary, miRDeep maps the sequenced small RNAs to the structures of candidate miRNA hairpins and scores the fit to a simple model of miRNA biogenesis. The score cutoff can be adjusted for trade-offs between sensitivity and specificity.

We found that a score cutoff of 3 recovered known nematode miRNAs present in the data (present meaning that miRBase *C. elegans* mature sequences that map perfectly to one or more excised potential precursors) with high sensitivity (80%), and the number of false positives was computationally estimated to be relatively low (7 ± 2).

Northern blots. Validation of 26G-RNA candidates was done by northern blot analysis as described previously[36]. We loaded 100 μg of total RNA from mixed embryos per lane. Probes used are listed in **Supplementary Table 9**. As 26G-RNAs were seemingly lowly expressed, the imaging plates had to be exposed for 4 h. Pictures were obtained with an imaging plate reader and processed in Adobe Illustrator.

RT-PCRs. RT-PCR was performed to examine the expression of early embryonic marker genes. The reverse transcritpion reaction was random primed. Primers for subsequent PCR are listed in **Supplementary Table 9**. We performed 35 cycles of PCR. Gel pictures were processed in Adobe Illustrator.

30. Stiernagle, T. Maintenance of *C. elegans*. *WormBook* **2006**, 1–11 (2006).
31. Hoffman, R.A. & Wood, J.C.S. Characterization of flow cytometer instrument sensitivity. *Curr. Protoc. Cytometry* **1.20**, 1.20.21–21.20.18 (2007).
32. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res.* **36**, D154–D158 (2008).
33. Jurka, J. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
34. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
35. Karolchik, D. *et al.* The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* **36**, D773–D779 (2008).
36. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853–858 (2001).

doi:10.1038/nmeth.1370

DISCUSSION

Discovery of miRNAs in deep sequencing data

The miRDeep model

The major challenge to identifying regulatory small RNAs in deep sequencing data is the abundances of degradation products from rRNAs, tRNAs, mRNAs or unknown sources that are also present in the data. We have built a model to identify known and novel miRNAs in deep sequencing data. The model implements current knowledge of miRNA biogenesis: the characteristic stable hairpin structure of the miRNA precursor; the Drosha/Dicer processing signature that causes miRNA products (mature, loop, star) to locate to stacks at particular positions in the hairpin precursors; and possible sequence conservation of the mature miRNA. To identify miRNA genes, we scan the genome, identifying loci that have deep sequencing reads mapping. Since we assume that these reads could be sequenced miRNA products, we fit the reads to the model. The fit to the model provides a log-odds score, reflecting the probability that a given locus is a genuine miRNA gene. More formally, the score is calculated from this equation:

$$Score = log(P(pre|data)/P(bgr|data)) \quad (1)$$

$$P(pre|data) = P(data|pre)P(pre)/P(data) \quad (2)$$
$$P(pre|data) - P(abs|pre)P(rel|pre)P(sig|pre)P(star|pre)P(nuc|pre)P(pre)/P(data) \quad (3)$$

$$P(bgr|data) = P(data|bgr)P(bgr)/P(data) \quad (4)$$
$$P(bgr|data) = P(abs|bgr)P(rel|bgr)P(sig|bgr)P(star|bgr)P(nuc|bgr)P(bgr)/P(data) \quad (5)$$

Where **P(pre)** is the probability that a given hairpin is a genuine miRNA precursor, **P(bgr)** is the probability that a given hairpin is a background non-miRNA hairpin, **abs** is the estimated absolute free energy (in kcal/mol) of the hairpin structure, **rel** equal to 1 if the potential precursor sequence is energetically stable, 0 otherwise, **sig** is the number of deep sequencing reads that are in consistency with Drosha/Dicer processing, **star** is equal to 0 if the potential precursor sequence has no reads that represent a putative star sequence, and 1 otherwise, **nuc** is equal to 1 if nucleotides 2–8 from the 5' end of the putative mature miRNA are not conserved in other

metazoans, and 1 otherwise. The | symbol denotes the Bayesian conditional probability and is read as 'given', e.g. **P (abs | pre)** is the probability that a hairpin would have the estimated free energy abs, given that it is a genuine miRNA precursor. All the parameters were estimated from *C. elegans* and *S. mediterranea* known miRNA hairpins and from random genomic hairpins unlikely to be genuine miRNA hairpins[110].

In these days, many classification tools are based on machine learning algorithms such as support vector machines. Machine learning algorithms are good choices when large amounts of training data are available and the designer of the algorithm does not know beforehand how important different biological features are to the classification. However, the disadvantage of machine learning is that the trained algorithm is like a 'black box' – it is difficult to say exactly how the algorithm classifies and hence what biological features are important. In choosing to implement miRDeep using simple Bayesian probabilistic statistics we have made a theoretically sound and transparent model. We further believe that the reason why miRDeep performs well across all metazoan clades tested is in part because of the relative simplicity of the model.

miRDeep controls

When a small RNA library is sequenced, the resulting deep sequencing reads will typically locate to millions of loci, most of which have no connection with miRNA biology. Likewise, metazoan genomes have millions of loci which are predicted to produce hairpin-like transcripts if transcribed (for instance, the human genome is predicted to contain ~11 million of such loci[149]). As can be imagined, the chance intersection of millions of reads with millions of hairpins will inevitable cause any model to produce some false positives. To estimate the number of false positives produced in the analysis of a given dataset, the miRDeep controls 'shuffle' the combinations of read signatures and structures, thus breaking any biological connections that might be between read signatures and structures. When the shuffled data is input to miRDeep, the number of predictions produced gives an estimate of the false positive rate.

miRDeep results (dog)

The domestic dog (*C. familiaris*) is increasingly being studied as a model for system for human disease such as cancer[150]. However, when we started our investigations, only six dog miRNAs were annotated in the public miRBase database[22]. From our collaborators we obtained dog lymphocyte total RNA, from which we prepared and deep sequenced a small RNA library. We analyzed the data with miRDeep, predicting more than 200 novel dog miRNA genes. If dogs have a number of miRNA genes comparable to other mammals, they will likely have 500-1000 miRNA genes in total. This shows that deep sequencing of a limited number of cell types can yield a

substantial fraction of all miRNAs in a complex animal. Further, a number of the miRNAs that we predicted are clear homologs to human miRNAs believed to be involved in disease (e.g. miR-17 - miR-20, miR-92, miR-142, miR-150, miR-155[94, 151-153]).

miRDeep results (nematode)

The nematode *C. elegans* was the first species in which miRNAs were detected[1, 14]. During the last ten years the nematode genome has been heavily mined for miRNAs using conventional cloning and Sanger sequencing[154], purely computational predictions[155], and deep sequencing analysis[50]. When we started our studies, 135 *C. elegans* miRNAs were annotated in miRBase[22]. In one study, we used miRDeep to predict 13 novel nematode miRNAs from already mined data obtained from deep sequencing of mixed-stage nematode populations[50, 110]. A number of these miRNAs were validated by northern blot analysis (4/5 tested). In a second study, Marlon Stoeckius developed a method to cleanly cell sort early nematode embryos (eFACS[140]). With this method he obtained clean samples of 1-cell stage embryos, 2-4-cell stage embryos, older embryos and mixed-stage embryos, from which small RNA libraries were prepared and deep sequenced. We analyzed the pooled data with miRDeep, predicting additional 19 novel miRNA genes[140]. These studies show that novel miRNAs can be discovered even in species that have already been heavily mined for miRNAs, given the correct combination of cell sorting, deep sequencing and prediction.

miRDeep results (planarian)

The planarian *S. mediterranea* is an emerging model system for regeneration and stem cell biology[156]. When we first started working with planarians, 63 miRNAs from this species had been deposited in miRBase[22]. We prepared and sequenced small RNA libraries from whole-body untreated planarians, irradiated planarians and from cell sorted neoblast stem cells. The pooled data was analyzed with miRDeep, yielding 61 novel miRNA genes[139]. A number of these were validated by northern blot or qPCR analysis (20/27 tested), and some were specifically upregulated in neoblasts and downregulated in irradiated planarians depleted in neoblasts. This suggests that these miRNAs (some of which were deeply conserved) may have a role in stem cell maintenance or function. Interestingly, the majority of the novel planarian miRNAs (34/61) displayed no sequence similarity to any known miRNAs, indicating that miRDeep can identify miRNA families that have evolved independently from those known from common model systems.

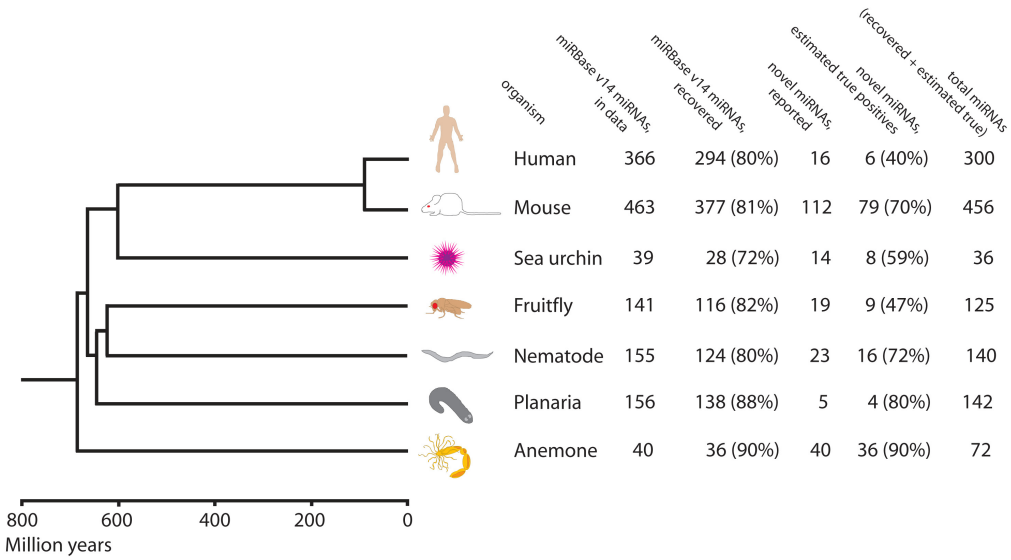


Figure 4. miRDeep reports miRNAs with consistent high sensitivity in seven animal clades. These seven clades were chosen for this figure to give good coverage of the metazoan phylogenetic tree. These particular analyses were in fact performed be the new version of miRDeep (see later section).

miRDeep results (other species)

Primarily as part of collaborations, we have now used miRDeep to predict miRNAs in around ten other species, including chimpanzee, pig, mouse, sea urchin, fruitfly and sea anemone (unpublished results). Taken together, these animals give good coverage of the animal phylogenetic tree. We find that miRDeep sensitivity and false positive rates are comparable between the species, even when the exact same parameters are used (see figure 4). This suggests that the miRDeep model captures features of miRNA biology that are shared between animals.

Prediction of non-canonical Drosha or Dicer hairpin substrates

In some of our miRDeep analyses, we have discarded reads that map to annotations of rRNAs, tRNAs or mRNAs. In the analysis of the deep sequenced dog small RNAs no reads were discarded because of annotation since little annotation was available. One of the novel dog miRNA hairpins predicted by miRDeep, miR-1306-5p, has since been shown to be homologous to a hairpin that is cleaved out of the human DGCR8 mRNA. The hairpin is cleaved by the Drosha endonuclease, reducing DGCR8 mRNA and protein levels[157]. This reduction does not depend on Dicer function. The hairpin is conserved in sequence and structure between human and dog, indicating that similar mechanisms lead to the production of the dog miR-1306-5p products. Analyzing mouse small RNA data, we have identified further five mRNAs that appear undergo similar processing (unpublished results), lending evidence to recent claims that Drosha cleavage might be a relatively widespread post-transcriptional regulatory mechanism[157, 158].

In a different study, we have analyzed deep sequencing reads from a library of small RNAs immunoprecipitated with Argonaute proteins in a human cell line[59]. In this study we also used miRDeep to predict miRNAs without first discarding reads that mapped to known annotations. Interestingly, one of the predicted miRNA hairpins located to the *bona fide* snoRNA ACA45. Extensive follow-up experiments performed by Christine Ender showed that the cleavage of ACA45 to the miRNA-like products is Dicer-dependent but Drosha-independent. Further, reporter assays indicated that the miRNA-like products can downregulate gene expression through 3'UTR base pairing[59].

These two examples show that miRDeep can recover Drosha/Dicer hairpin substrates that are not canonical miRNAs. The examples also raise the interesting question if miRDeep primarily detects Drosha or Dicer hairpin substrates. The cleavage of the DGCR8 hairpin is Drosha-dependent but Dicer-independent. Reversely, the cleavage of ACA45 is Dicer-dependent but Drosha-independent. Probably the question is ill posed, since Drosha and Dicer have likely co-evolved to bind to the same hairpin structures, given that the majority of Drosha products undergo downstream cleavage by Dicer.

Profiling miRNA expression using deep sequencing data

Limitations: absolute quantitation

While profiling small RNAs in planarians, we obtained three deep sequenced small RNA libraries from whole-body planarians, sequenced independently with the three major deep sequencing platforms: 454 / Life sciences, Solexa / Illumina and ABI SOLiD. To estimate if the platforms correlate in quantitating miRNA expression, we compared the normalized read counts for individual miRNAs between the three platforms (normalization as in Friedländer *et al.*[139]). We found only weak correlation (Pearson's correlation < 0.5 in all pairwise comparisons, data not shown). Similarly, we compared the miRNA read counts with miRNA expression as roughly estimated from the qPCR RT values, and again found only weak correlations (Pearson's correlation < 0.5). Our results suggest that none of the deep sequencing platforms are capable of precisely quantitating absolute miRNA expression. These findings have recently been independently validated by systematic studies using stoichiometrically controlled synthetic miRNAs[159]. The same study shows that the biases are primarily the result of ligation biases in the library preparation.

Possibilities: fold-changes

We compared Solexa miRNA read counts between two biological replicates of planarian samples and found excellent correlation (Pearson's correlation > 0.99) showing that at least Solexa miRNA deep sequencing is highly reproducible[139]. In addition, we found that miRNA expression fold-changes, as quantitated by Solexa sequencing, correlated well with expression fold-changes, as quantitated by qPCR (Pearson's correlation = 0.93)[139]. This indicates that at least Solexa sequencing can be used to precisely profile changes in miRNA expression across two samples. These results have recently been expanded to also 454 and SOLiD sequencing[159].

'Blind spots'

When comparing miRNA read counts across the three platforms, we found a (low) number of miRNAs that were sequenced many times by two of the platforms, but only very few times by the third of the platforms. This suggests that the three platforms each have a (low) number of miRNAs that they have difficulties detecting. Further insights into this will likely require systematic chemical studies of biases in the ligation protocols used by the three platform.

piRNA deep sequencing analysis

Identifying piRNA populations

It was the deep sequencing technology that first allowed investigation into the biology of the abundant piRNAs in mouse and fly. In most studies, immunoprecipitation of Piwi proteins is used to isolate distinct piRNA populations. Since we did not have available antibodies for the planarian Piwi proteins, we obtained deep sequenced libraries of total cellular small RNAs, and computationally separated miRNAs and degradation products of mRNA, rRNA and tRNAs. The remaining ~1.2 million deep sequencing reads had distinct length peak at ~32 nts, and as a population displayed similar piRNA sequence biases and genome clustering patterns as have been observed in other species. We even found that the reads mapping sense to transposable elements had distinct sequence biases as the reads that mapped antisense, suggesting that they might associate with distinct Piwi proteins. The fact that the planarian primary piRNAs, like the fly ones, map antisense to transposable elements suggests that this form of ping-pong loop is ancestral, while the mouse 'upside-down' ping-pong loop is likely derived. We also established a method to estimate total piRNA abundances in distinct samples by normalizing total piRNA read counts to the read counts of individual miRNAs that we observe are constantly expressed across the samples according to qPCR assays. In sum, our studies indicate that it is possible to identify and profile

piRNA populations and subpopulations in part without using immunoprecipitation, given the correct computational analysis.

piRNAs in stem cells and in the germ line

Planarian neoblast stem cells are the only cells not associated with the germline in which piRNAs have been observed. What is the function that piRNAs perform in stem cells and the germ line, and why are they not observed in other cells? It is currently believed that the primary function of piRNAs is to silence transposable elements[127, 160]. These are mainly active in the germ line consistent with the fact that only germ line cells are perpetuated to the next generations, while somatic cells die with each individual animal. Similarly, it is most important for the host to silence transposons in the germ line, where they are most active and where new copies of transposable elements can be perpetuated to the next generations. In asexual planarians, the neoblast stem cells are the cells that are perpetuated and thus important to protect from transposon activity. When the Piwi2 gene is knocked down in planarians, the neoblasts can undergo mitosis but are unable to terminally differentiate into somatic cells. This depletes the pool of neoblasts, eventually causing death of the planarian[161]. This phenotype is consistent with the role of Piwi proteins in silencing transposons. In Piwi2 knockdown planarians, transposons would actively re-insert into new genomic positions in neoblasts, causing widespread DNA damage. However, unlike the mouse and fly germ cells which arrest in meiosis, the planarian neoblasts would likely still be able to complete mitosis in spite of the DNA damage, given that the mitotic check-point control is generally less stringent than the meiotic check-point (personal communication, Alexei Aravin). However, the neoblasts might not be able to undergo terminal differentiation, given the genetic damage they have sustained. A simple way to test this hypothesis would be to do *in situ* hybridization of Piwi2 knockdown planarians with antibodies against protein markers of DNA damage. Further, our lab is currently purifying antibodies against the Piwi2 protein. This will hopefully allow us to perform CLIP-seq and CHIP-seq to identify the RNA and DNA targets of the Piwi2 interacting small RNAs.

Possible improvements

miRDeep2

Improved work flow

Since the first version of miRDeep was published, we have designed and implemented a new version (manuscript in preparation). miRDeep2 not only discovers known and novel miRNAs in deep sequencing data, but also includes a module that can process raw sequencing reads and map them to the reference genome (Mapper) and a module that can perform fast and exact quantitation of known miRNAs in a given dataset (Quantifier). The modules work complementary, e.g. output of Mapper can be directly input to the miRDeep2 module or the Quantifier module (figure 5).

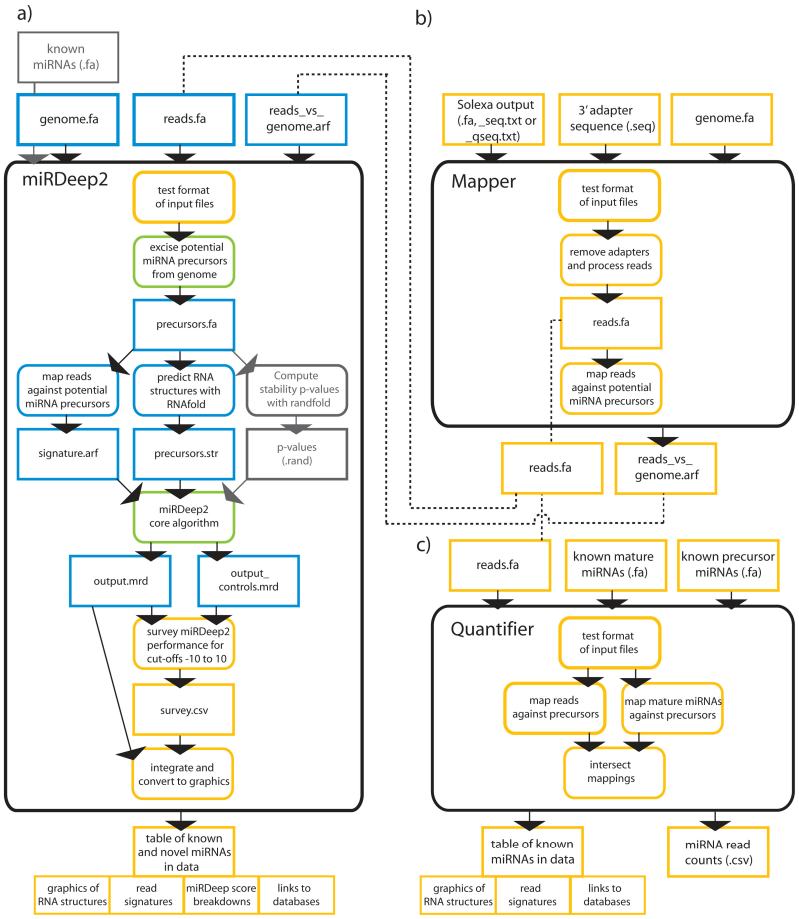


Figure 5. Flow charts for (a) the miRDeep2 module (identifies known and novel miRNAs in deep sequencing data), (b) the Mapper module (processes Solexa / Illumina output and maps it to the reference genome) and (c) the Quantifier module (sums up miRNA read counts in a deep sequencing set). For each module the input, internal work flow (in black borders) and output is shown. Files are presented in rectangular boxes; processes are presented in rounded boxes. Files and processes that are novel to miRDeep2 are in yellow, those that have been modified are in green and those that remain largely unchanged are in blue. Files and processes that are optional are in grey.

Reduced memory and time consumption

Given that the number of reads output by the deep sequencing platforms have increased by a factor ten every two years, and given that this trend may continue, it is essential that tools that analyze the data are efficient in terms of memory and time consumption. While the first version of miRDeep could require weeks of computing on high-memory (256 GB memory) computers to process, map and classify ~50 million reads, miRDeep2 can perform the same analysis in six hours on a desktop computer with 4 GB memory. We project that run-time doubles for each ten-fold increase in read input. Further, miRDeep2 output consists of a webpage table of all known and novel miRNAs in the data. This webpage links to graphical representation of all results produced (such as miRNA hairpin structures, read signatures, score break-downs, see figure 6) as well as links to public databases such as miRBase, the UCSC genome browser, NCBI blast search etc.

Robust analysis of very deep data

More importantly, the new version can analyze very deep sequencing data in a more robust manner, without getting distracted by non-canonical Drosha/Dicer products. When deep sequenced miRNAs are mapped back to their genome locus, they typically map in three piles, corresponding to mature, loop and star sequences. When the sequencing is deep enough, reads sometimes map in adjacent piles (see figure 6). These reads are likely degradation products of the primary transcript from which the miRNA hairpin was cut out of. In the first version of miRDeep, the boundary of the putative miRNA precursor on the genome was defined by the local *cluster* of reads, meaning that the adjacent piles of degradation products would be included in the precursor, causing a misrepresentation of the boundary. In miRDeep2, the boundary of the putative miRNA precursor is defined from the single highest local *stack* of reads, which are assumed to represent sequenced mature miRNAs. The boundary of the miRNA precursor is then defined as the genome sequence covered by the stack, plus flanking sequence corresponding to typical miRNA loop and star length. Since the excision is based only on the stack of reads from the mature miRNA, the algorithm is insensitive to adjacent stacks that can potentially disturb identification of the miRNA hairpin.

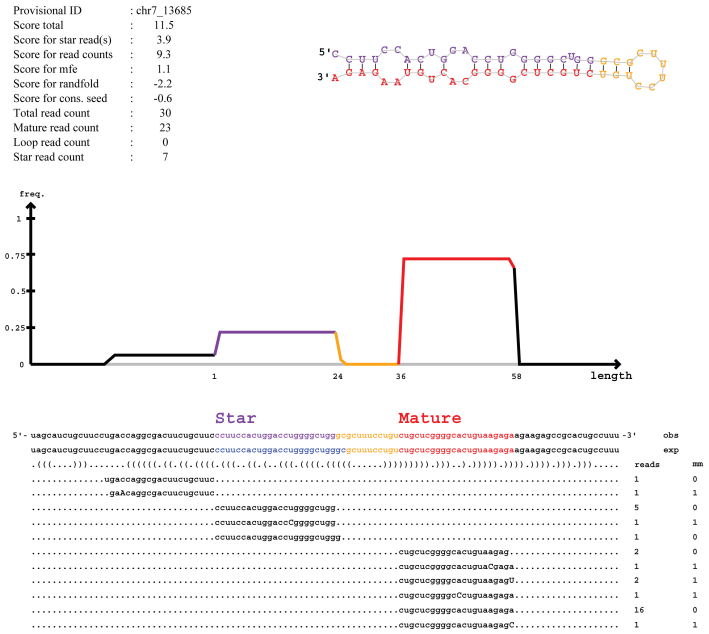


Figure 6. Novel mouse miRNA predicted by miRDeep2. Notice the two reads that locate immediately upstream of the miRNA precursors. This is an example of miRDeep2 graphic output.

Integrated annotation

It is trivially true that if we could correctly identify all non-miRNA reads in a deep sequencing dataset, we would know that the remaining were miRNA reads. But in our studies we have come across a number of examples where initial identification of non-miRNA reads in practice improves miRNA prediction:

Integrated annotation of 21U-RNAs and miRNAs

21U-RNAs and miRNAs both often have beginning uracils and length of ~22 nts. This can make them difficult to distinguish computationally. In some cases, 21U-RNAs locate to genomic hairpins and sometimes two 21U-RNAs locate to hairpins in positions such that they resemble mature and star miRNA products. However, the characteristic upstream motif makes 21U-RNAs easy to identify. In the nematode embryo study[140], we found that it was difficult to identify novel miRNAs with confidence when miRDeep was run on the unfiltered data. However, prediction accuracy improved substantially when the upstream motif was first used to identify and filter out 21U-RNAs.

Integrated annotation of piRNAs and miRNAs

The planarian piRNAs are typically ~32 nts in lengths and are in most cases easy to distinguish from miRNAs. However, shorter and possible partly degraded piRNAs are sometimes sequenced. For instance, if a given piRNA is present full-length in the data in ten copies, there might be a single 22 nt copy as well. If only reads in the canonical miRNA length range (18-25 nts) are input to miRDeep, these short piRNAs can cause large numbers of false positives. Therefore we make sure to input reads of all length. If the consensus putative mature miRNA is longer than 26 nts, miRDeep will automatically flag it as a possible piRNA and will discard the candidate gene.

'Black matter' of sequenced short RNAs

These cases show that improved identification of one type of small RNAs will also improve identification of other sources of small RNAs. Currently, it is not well understood exactly what are the sources of the sequenced short RNAs that we observe in the deep sequencing data. It is assumed that there are a number of degradation products from rRNAs, tRNAs, mRNAs etc. along with the regulatory small RNAs. But recent studies indicate that even some of the 'degradation products' show regularities in terms of their length and positions in the longer transcripts[162]. Thus, also rRNAs, tRNAs, scRNAs etc. might be substrates for processing activity. More importantly, it is rarely possible to assign more than 90% of all deep sequencing reads to known transcripts. Currently we simply have no idea what are the sources of the remaining hundreds of thousands of reads. Eventually, as more biological knowledge of the sources of the short RNAs become available, it will hopefully be possible to make models to make an integrated and saturated identification of all short RNAs in any given deep sequencing dataset.

Mapping one read to one locus

We know from biology that each deep sequenced RNA has been transcribed from exactly one genome locus. However, when sequenced small RNAs are mapped to the reference genome, many map to more than one locus. This is in some cases because the RNA is transcribed from a gene with many copies in the genome, like a transposable element. In some cases it will be 'spurious' mappings, meaning that a short sequence can have chance matches to biologically unrelated positions in the genome, especially when the reference genome is large.

Discarding ambiguous mappers

One method of resolving ambiguously mapping reads is to discard them. In this way, it is fairly certain that all reads that undergo downstream analysis have been traced to the correct genome sources. However, this method discards a lot of useful information. For instance, most piRNAs and many miRNAs map to more than locus in the genome.

Retaining all mappings

Another method is to retain all equally best mappings of each read. This is the method that has been used in the present studies. To avoid that reads with many mappings dominate the analysis, all read mappings can be assigned weights that are the inverse to the number of times the read maps (e.g. each mapping of a read that maps ten times will be assigned a weight of 0.1). The advantage of this method is that little information is thrown away. The disadvantages are a) we know that the solution is incorrect since a given read can only have one genomic source b) the large numbers of mappings make the downstream analysis more noisy.

'Parsimonious mapping'

A solution to the problem could be to assume that most deep sequencing reads have originated from a relatively small number of genome loci, and attempt to map the reads such that most of them locate to the fewest possible number of loci. In some concrete cases this appears reasonable. For instance, imagine a read that maps equally well to two genome loci. One locus is a 'read desert' with no other reads mapping nearby. The other locus is an rRNA gene that has thousands of reads mapping. In this case, it would seem reasonable to assume that the read should be mapped to the rRNA locus. A 'parsimonious' mapping method could be implemented in a number of ways. One way would be first to identify all reads that can be traced to a single genome loci. These would serve as a mapping 'scaffold'. Then the ambiguous mappers would be assigned to the positions that maximize the 'parsimony' of the mappings, understood as the mappings that would assign most of the reads to the fewest loci (by some clearly defined criteria). The assignment of the ambiguous mappers could be resolved with Monte Carlo simulations[163].

The future of small RNA deep sequencing

This last section will in a speculative manner explore the possibilities that open up if the emerging deep sequencing platforms will be able to increase read output by two orders of magnitude (hundred fold increase). This does not seem unlikely given that this is the increase that has been observed the last four years. The section will primarily discuss how the technology might impact miRNA studies, but the arguments can in most cases be extrapolated to other types of studies.

More depth, more samples:

One possible use of a hundred fold increase in sequencing depth is naturally to sequence hundred samples simultaneously. A 'sample' is here defined as collection of cells from a specific tissue and/or developmental stage from a given organism. It is assumed here that the samples have been properly prepared for sequencing (the RNA has been extracted, small RNA libraries prepared etc.). The existing deep sequencing platforms (454, Solexa, SOLiD) already include protocols for the simultaneous sequencing of multiple samples ('multiplexing'). This is done in the following way: the sample libraries are prepared in series and are ligated with adapters with distinct sequences (barcodes). Then the samples are pooled and sequenced simultaneously. When the sequences are analyzed, the barcode on each read allows it to be traced back to the source sample. If the output of the deep sequencing platforms increases by two orders of magnitude, multiplexing will open up these possibilities:

Expression profiling

There has already been a study to make an atlas of miRNA expression in the human body[32]. However, this was a vast project given that it was undertaken with conventional cloning and Sanger sequencing. If the output of deep sequencing increases substantially, it might be possible to make an atlas of miRNA expression in hundreds of tissues in an organism in a single sequencing.

Discovery

Discovery of novel miRNAs is often limited by the number of miRNAs that are expressed in the tissue that is investigated. Typically a dozen miRNAs will be highly expressed (and thus easy to discover) in a given tissue, while hundreds of miRNAs will be very lowly expressed (and thus difficult to discover). If hundreds of tissues are sequenced simultaneously, it would be expected that most functional miRNAs are highly expressed in at least one tissue, thus making discovery of

all miRNAs easy. Alternatively, a single tissue could be interrogated across dozens of animal species. In this case, putative miRNAs that are very lowly expressed in all species but display some sequence conservation might still be confidently identified as miRNAs.

Independent validation

The current small RNA library preparation steps all include a PCR amplification step. This means that a single small RNA molecule in a sample will be present in several copies in the amplified library, and could also be sequenced several times. This makes it difficult for the computational biologist to interpret the data: does a given sequence occur multiple times in the data because it occurred multiple times in the sample, or because the library was PCR amplified? The distinction can be important, since a sequence that occurs multiple times in the sample can be taken as evidence that biogenesis lead to the accumulation of the specific sequence. In contrast, a sequence that occurs a single time in the sample is more likely a product of degradation. If, however, a given sequence occurs in two distinct samples, then the occurrences can truly be considered as independent evidence (to my knowledge, this problem has not been described in the literature).

Limitations to multiplexing

The practical limitations to the simultaneous sequencing of multiple samples have already been mentioned. The tissues have to be isolated, or cells harvested, in series. Also, with the current sequencing platforms the library preparation is a significant bottleneck. Since multiplexed libraries have to be individually fitted with barcodes they have to be prepared in series, which means that little work is saved. However, these problems might be alleviated in the future as protocols improve.

More depth, one sample:

As an alternative to multiplexing, increased sequencing depth can be used to sequence a single sample to saturation.

Sequencing to profile genome-wide degradation and small RNA expression

A two order of magnitude increase in sequencing depth could lead to qualitative as well as quantitative improvements in the data. For instance, if the small RNA contents of a human cell line were sequenced to produce 10 billion reads of up to 50 nucleotides length, this would mean that every nucleotide in the reference human genome would be covered by more than 15 reads on average (although the reads would not distribute evenly). The number of reads covering each nucleotide could then be taken as a measure of expression of small RNAs and of transcripts being degraded to short RNAs. The data would much resemble that produced by a genome tiling array[164] in that it would show expression at high resolution genome-wide.

Comparison to genome tiling arrays

Deep sequencing data would however have advantages over the genome tiling array: a) tiling arrays only have single probe (expression readout) for every 20 nucleotide or so of the genome. In comparison, the deep sequencing data would have an expression readout for every nucleotide of the genome. b) tiling arrays do not confer any information on the length of the transcripts bound to each probe. The length of each read gives this information in the deep sequencing data. c) in tiling arrays it is assumed that the transcripts bound to each probe has (reverse complement) similarity to the probe. However, cross-hybridization remains an issue. In comparison, deep sequencing provides the exact sequence of each transcript.

Distinguishing degradation products and regulatory small RNAs

One use of such saturated small RNA deep sequencing would be that it is easy to distinguish degradation products from regulatory small RNAs genome-wide. If sequenced RNAs mapping to a genome locus have varying lengths and have offset begin positions (like fallen dominoes) then the RNAs are likely degradation products of longer transcripts. This information is also essentially useful. If the sequenced RNAs mapping to a locus have specific lengths and locate to stacks, then they are likely cleavage products of small RNA biogenesis. For instance, a miRNA locus would likely be characterized by three stacks of reads, with the two outer stacks consisting of reads ~22 nucleotides in length; with one of the outer stacks consisting of the most reads (sequenced mature miRNAs) and the middle stack consisting of the fewest reads (sequenced miRNA loops). The

genome surrounding these stacks might have many reads mapping from degradation of the miRNA primary transcript (which is visible in genome tiling arrays[158]).

Discovering miRNAs by counting read stacks

As mentioned, the increasingly deep sequencing has demanded increasingly sophisticated algorithms to discern the few miRNA loci from the millions of non-miRNA loci analyzed. However, as the deep sequencing platforms improve, this trend might be reversed. If sequencing of small RNAs gets sufficiently saturated, miRNA loci might be confidently identified through simple stack counting, as described above. In this case, simple algorithms might outperform the sophisticated algorithms like miRDeep.

From identification to function

Ultimately, the reason why we want to identify novel small RNAs is that we want to know their function. We want to know how they influence human disease, development, evolution etc. Function can in part be unravelled through computational predictions (like miRNA target prediction) or through novel high-throughput technologies like CLIP-seq, CHIP-seq, RNA-seq, mass spectrometry or ultimately through analysis of knockout phenotypes. Such experiments will likely suggest that many of small RNAs that have been discovered with the help of sensitive deep sequencing do not have any discernable function. It is completely plausible that hundreds of say human hairpin transcripts undergo Drosha/Dicer cleavage and may even be incorporated into the miRNP effector complex, but are expressed at such low levels that they have no real impact on protein output in any cellular context. Such very lowly expressed Drosha/Dicer substrates might be tolerated because they have little detrimental effect, and may during evolution eventually be selected against or may develop some function[165]. As the sequencing gets deeper, it is likely that more of such very lowly expressed Drosha/Dicer substrates will be discovered. Thus in the future, while it will be technically and computationally easier to identify novel small RNAs, the field will turn in a more philosophical direction: how much function does a small RNA need to have in order to be a genuine regulatory small RNA?

AUTHOR CONTRIBUTIONS OF THE DOCTORATE STUDENT

Friedländer *et al.*, 'Discovering microRNAs from deep sequencing data using miRDeep', Nature Biotechnology (2008)

Marc Riemer Friedländer (MRF) designed and developed miRDeep and performed all computational analysis described in this study. This was done under supervision by Nikolaus Rajewsky (NR). MRF and NR wrote the manuscript.

Friedländer *et al.*, 'High-resolution profiling and discovery of planarian small RNAs', PNAS (2009)

MRF developed all computational tools used in this study (except where otherwise explicitly stated) and performed all computational analysis. This was done under supervision by NR. MRF, NR and Catherine Adamidi wrote the manuscript.

Ender *et al.*, 'A Human snoRNA with MicroRNA-Like Functions', Molecular Cell, (2008)

MRF first identified ACA45 as a Dicer product using miRDeep and performed the analysis showing that processing of the snoRNA is conserved in a number of mammals. This was done under supervision by NR. MRF made figure 2B and wrote the methods sections corresponding to the analysis performed.

Stoeckius *et al.*, 'Large-scale sorting of *C. elegans* embryos reveals the dynamics of small RNA expression', Nature Methods, (2009).

MRF identified 19 novel nematode miRNAs. This was done under supervision by NR. MRF wrote the results and methods sections corresponding to the analysis performed.

SUMMARY IN ENGLISH AND GERMAN

Discoveries in the last decade have shown that small RNAs (such as microRNAs) perform a number of important functions, including post-transcriptional gene regulation, transposon silencing, DNA methylation, chromatin modifications and chromosome segregation. The ability of the new deep sequencing technologies to sequence millions of short RNAs in a few hours have made them the method of choice for simultaneous discovery and profiling of small RNAs. However, when the sequenced RNAs are mapped to the reference genome, they typically locate to millions of distinct loci, only a few of which are loci that produce regulatory small RNAs. To distinguish the few loci that produce regulatory small RNAs from the many loci that are sources of other short RNAs like degradation products is a non-trivial computational challenge. In my doctorate works I have formalized knowledge of small RNA biology and biogenesis into computational models that can accurately identify regulatory small RNAs of different classes in much larger pools of sequenced RNAs. As part of collaborations, I have used these models to discover hundreds of novel small RNA genes in more than ten animal species including humans, mice, fruit flies, nematodes and planarian flatworms. We find evidence that a number of these small RNA genes have roles in disease or in stem cell function. Further, some of the novel regulatory small RNAs are in fact cleaved *bona fide* snoRNAs, revealing cross-talk between two RNA pathways. Last, I have developed methods for precise quantitation of individual small RNAs as well as entire small RNA populations between deep sequencing samples.

Die Entdeckungen des letzten Jahrzehnts haben gezeigt, dass so genannte *small RNAs*, wie zum Beispiel microRNAs, bedeutenden Einfluss auf viele Zellabläufe haben. Dazu zählen unter anderem posttranskriptionelle Regulation, Chromatin Modifikationen sowie Segregation der Chromosomen. So genannte next generation sequencer Maschinen sind dazu in der Lage Millionen von kurzen RNA Molekülen innerhalb nur werniger Stunden zu sequenzieren, weshalb sie heutzutage das Mittel Wahl sind um sowohl neue regulatorische small RNAs zu entdecken als auch Expressionsprofile von diesen zu erstellen. Wenn die sequenzierten RNAs auf das Genom gemappt werden gibt es normalerweise Millionen von verschieden Moeglichkeiten von dem sie stammen koennten, aber nur einige von Ihnen produzieren kleine regulatorische RNAs.
Die Identifikation genau dieser wenigen Loci, von denen die kleinen regulatorischen RNA Stücke stammen, ist eine computertechnisch anspruchsvolle Aufgabe. In meiner Doktorarbeit habe ich computerbasierte Modelle auf der Grundlage von der Biologie und Biogenese kleiner RNAs erstellt. Diese Modelle sind dazu in der Lage die Loci der verschiedenen kleinen regulatorischen RNAs zuverlaessig zu identifizieren. Während diverser Kollaborationen mit anderen

Arbeitsgruppe habe ich meine Modelle dazu benutzt hunderte von noch nicht detektierten kleinen RNA Genen in mehr als zehn verschiedenen Tierspezies zu identifizieren. Dazu zählen Menschen, Mäuse, Fruchtfliegen und Flachwürmer. Wir haben Evidenz dafür gefunden, dass eine Vielzahl dieser kleinen RNAs eine Rolle in diversen Krankheiten oder Stammzellfunktion spielen. Des Weiteren sind einiger dieser kleinen RNAs tatsächlich prozessierte snoRNAs sind, was auf eine Interaktion der verschiedenen RNA Pathways nahelegt. Als Letzes habe ich noch Methoden entwickelt, die eine präzise Quantifikation von einzelnen kleinen RNAs sowie gesamter Populationen von kleinen RNAs zwischen Proben erlauben.

REFERENCES

1. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855-862 (1993).
2. Aravin, A.A., Sachidanandam, R., Girard, A., Fejes-Toth, K. & Hannon, G.J. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316**, 744-747 (2007).
3. Verdel, A. et al. RNAi-mediated targeting of heterochromatin by the RITS complex. *Science* **303**, 672-676 (2004).
4. Claycomb, J.M. et al. The Argonaute CSR-1 and its 22G-RNA cofactors are required for holocentric chromosome segregation. *Cell* **139**, 123-134 (2009).
5. Farazi, T.A., Juranek, S.A. & Tuschl, T. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development* **135**, 1201-1214 (2008).
6. Ghildiyal, M. & Zamore, P.D. Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10**, 94-108 (2009).
7. Berninger, P., Gaidatzis, D., van Nimwegen, E. & Zavolan, M. Computational analysis of small RNA cloning data. *Methods* **44**, 13-21 (2008).
8. Hafner, M. et al. Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing. *Methods* **44**, 3-12 (2008).
9. Barrangou, R. et al. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**, 1709-1712 (2007).
10. Brouns, S.J. et al. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* **321**, 960-964 (2008).
11. Cerutti, H. & Casas-Mollano, J.A. On the origin and functions of RNA-mediated silencing: from protists to man. *Curr Genet* **50**, 81-99 (2006).
12. Chapman, E.J. & Carrington, J.C. Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet* **8**, 884-896 (2007).
13. Millar, A.A. & Waterhouse, P.M. Plant and animal microRNAs: similarities and differences. *Funct Integr Genomics* **5**, 129-135 (2005).
14. Lee, R.C., Feinbaum, R.L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854 (1993).
15. Hammond, S.M., Bernstein, E., Beach, D. & Hannon, G.J. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature* **404**, 293-296 (2000).
16. Olsen, P.H. & Ambros, V. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol* **216**, 671-680 (1999).
17. Wu, L., Fan, J. & Belasco, J.G. MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci U S A* **103**, 4034-4039 (2006).
18. Bagga, S. et al. Regulation by let-7 and lin-4 miRNAs results in target mRNA degradation. *Cell* **122**, 553-563 (2005).
19. Behm-Ansmant, I. et al. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev* **20**, 1885-1898 (2006).
20. Giraldez, A.J. et al. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. *Science* **312**, 75-79 (2006).
21. Grimson, A. et al. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**, 1193-1197 (2008).
22. Griffiths-Jones, S., Saini, H.K., van Dongen, S. & Enright, A.J. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36**, D154-158 (2008).
23. Pasquinelli, A.E. et al. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86-89 (2000).
24. Lee, R.C. & Ambros, V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**, 862-864 (2001).
25. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853-858 (2001).

26. Lau, N.C., Lim, L.P., Weinstein, E.G. & Bartel, D.P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858-862 (2001).
27. Krek, A. et al. Combinatorial microRNA target predictions. *Nat Genet* **37**, 495-500 (2005).
28. Friedman, R.C., Farh, K.K., Burge, C.B. & Bartel, D.P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* **19**, 92-105 (2009).
29. Lim, L.P. et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769-773 (2005).
30. Kloosterman, W.P. & Plasterk, R.H. The diverse functions of microRNAs in animal development and disease. *Dev Cell* **11**, 441-450 (2006).
31. Bushati, N. & Cohen, S.M. microRNA functions. *Annu Rev Cell Dev Biol* **23**, 175-205 (2007).
32. Landgraf, P. et al. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401-1414 (2007).
33. Chalfie, M., Horvitz, H.R. & Sulston, J.E. Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell* **24**, 59-69 (1981).
34. Georges, M. et al. Polymorphic microRNA-target interactions: a novel source of phenotypic variation. *Cold Spring Harb Symp Quant Biol* **71**, 343-350 (2006).
35. Lee, Y. et al. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J* **23**, 4051-4060 (2004).
36. Cai, X., Hagedorn, C.H. & Cullen, B.R. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA* **10**, 1957-1966 (2004).
37. Lee, Y., Jeon, K., Lee, J.T., Kim, S. & Kim, V.N. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J* **21**, 4663-4670 (2002).
38. Denli, A.M., Tops, B.B., Plasterk, R.H., Ketting, R.F. & Hannon, G.J. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**, 231-235 (2004).
39. Gregory, R.I. et al. The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432**, 235-240 (2004).
40. Han, J. et al. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**, 887-901 (2006).
41. Zeng, Y. & Cullen, B.R. Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences. *J Biol Chem* **280**, 27595-27603 (2005).
42. Lund, E., Guttinger, S., Calado, A., Dahlberg, J.E. & Kutay, U. Nuclear export of microRNA precursors. *Science* **303**, 95-98 (2004).
43. Bohnsack, M.T., Czaplinski, K. & Gorlich, D. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *RNA* **10**, 185-191 (2004).
44. Bernstein, E., Caudy, A.A., Hammond, S.M. & Hannon, G.J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363-366 (2001).
45. Grishok, A. et al. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell* **106**, 23-34 (2001).
46. Hutvagner, G. et al. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834-838 (2001).
47. Ketting, R.F. et al. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev* **15**, 2654-2659 (2001).
48. Knight, S.W. & Bass, B.L. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science* **293**, 2269-2271 (2001).
49. Bonnet, E., Wuyts, J., Rouze, P. & Van de Peer, Y. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics* **20**, 2911-2917 (2004).
50. Ruby, J.G. et al. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell* **127**, 1193-1207 (2006).
51. Khvorova, A., Reynolds, A. & Jayasena, S.D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209-216 (2003).
52. Schwarz, D.S. et al. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199-208 (2003).
53. Okamura, K. et al. The regulatory activity of microRNA* species has substantial influence on microRNA and 3' UTR evolution. *Nat Struct Mol Biol* **15**, 354-363 (2008).
54. Ruby, J.G. et al. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res* **17**, 1850-1864 (2007).
55. Rodriguez, A., Griffiths-Jones, S., Ashurst, J.L. & Bradley, A. Identification of mammalian microRNA host genes and transcription units. *Genome Res* **14**, 1902-1910 (2004).

56. Ruby, J.G., Jan, C.H. & Bartel, D.P. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**, 83-86 (2007).
57. Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M. & Lai, E.C. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* **130**, 89-100 (2007).
58. Berezikov, E., Chung, W.J., Willis, J., Cuppen, E. & Lai, E.C. Mammalian mirtron genes. *Mol Cell* **28**, 328-336 (2007).
59. Ender, C. et al. A human snoRNA with microRNA-like functions. *Mol Cell* **32**, 519-528 (2008).
60. Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B. & Bartel, D.P. MicroRNAs in plants. *Genes Dev* **16**, 1616-1626 (2002).
61. Lai, E.C. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* **30**, 363-364 (2002).
62. Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350-355 (2004).
63. Selbach, M. et al. Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58-63 (2008).
64. Baek, D. et al. The impact of microRNAs on protein output. *Nature* **455**, 64-71 (2008).
65. Vinther, J., Hedegaard, M.M., Gardner, P.P., Andersen, J.S. & Arctander, P. Identification of miRNA targets with stable isotope labeling by amino acids in cell culture. *Nucleic Acids Res* **34**, e107 (2006).
66. Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P. & Burge, C.B. Prediction of mammalian microRNA targets. *Cell* **115**, 787-798 (2003).
67. Lewis, B.P., Burge, C.B. & Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20 (2005).
68. Chi, S.W., Zang, J.B., Mele, A. & Darnell, R.B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* **460**, 479-486 (2009).
69. Licatalosi, D.D. et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464-469 (2008).
70. Yeo, G.W. et al. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nat Struct Mol Biol* **16**, 130-137 (2009).
71. Sanford, J.R. et al. Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts. *Genome Res* **19**, 381-394 (2009).
72. Filipowicz, W., Bhattacharyya, S.N. & Sonenberg, N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet* **9**, 102-114 (2008).
73. Kiriakidou, M. et al. An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* **129**, 1141-1151 (2007).
74. Chendrimada, T.P. et al. MicroRNA silencing through RISC recruitment of eIF6. *Nature* **447**, 823-828 (2007).
75. Petersen, C.P., Bordeleau, M.E., Pelletier, J. & Sharp, P.A. Short RNAs repress translation after initiation in mammalian cells. *Mol Cell* **21**, 533-542 (2006).
76. Nottrott, S., Simard, M.J. & Richter, J.D. Human let-7a miRNA blocks protein production on actively translating polyribosomes. *Nat Struct Mol Biol* **13**, 1108-1114 (2006).
77. Maroney, P.A., Yu, Y., Fisher, J. & Nilsen, T.W. Evidence that microRNAs are associated with translating messenger RNAs in human cells. *Nat Struct Mol Biol* **13**, 1102-1107 (2006).
78. Zipprich, J.T., Bhattacharyya, S., Mathys, H. & Filipowicz, W. Importance of the C-terminal domain of the human GW182 protein TNRC6C for translational repression. *RNA* **15**, 781-793 (2009).
79. Fabian, M.R. et al. Mammalian miRNA RISC recruits CAF1 and PABP to affect PABP-dependent deadenylation. *Mol Cell* **35**, 868-880 (2009).
80. Schmitter, D. et al. Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res* **34**, 4801-4815 (2006).
81. Aleman, L.M., Doench, J. & Sharp, P.A. Comparison of siRNA-induced off-target RNA and protein effects. *RNA* **13**, 385-395 (2007).
82. Sheth, U. & Parker, R. Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science* **300**, 805-808 (2003).
83. Teixeira, D., Sheth, U., Valencia-Sanchez, M.A., Brengues, M. & Parker, R. Processing bodies require RNA for assembly and contain nontranslating mRNAs. *RNA* **11**, 371-382 (2005).
84. Balagopal, V. & Parker, R. Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs. *Curr Opin Cell Biol* **21**, 403-408 (2009).

85. Bhattacharyya, S.N., Habermacher, R., Martine, U., Closs, E.I. & Filipowicz, W. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell* **125**, 1111-1124 (2006).
86. Pillai, R.S. et al. Inhibition of translational initiation by Let-7 MicroRNA in human cells. *Science* **309**, 1573-1576 (2005).
87. Liu, J., Valencia-Sanchez, M.A., Hannon, G.J. & Parker, R. MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nat Cell Biol* **7**, 719-723 (2005).
88. Leung, A.K., Calabrese, J.M. & Sharp, P.A. Quantitative analysis of Argonaute protein reveals microRNA-dependent localization to stress granules. *Proc Natl Acad Sci U S A* **103**, 18125-18130 (2006).
89. Eulalio, A., Behm-Ansmant, I., Schweizer, D. & Izaurralde, E. P-body formation is a consequence, not the cause, of RNA-mediated gene silencing. *Mol Cell Biol* **27**, 3970-3981 (2007).
90. Pauley, K.M. et al. Formation of GW bodies is a consequence of microRNA genesis. *EMBO Rep* **7**, 904-910 (2006).
91. Chu, C.Y. & Rana, T.M. Translation repression in human cells by microRNA-induced gene silencing requires RCK/p54. *PLoS Biol* **4**, e210 (2006).
92. Ruvkun, G. & Giusto, J. The Caenorhabditis elegans heterochronic gene lin-14 encodes a nuclear protein that forms a temporal developmental switch. *Nature* **338**, 313-319 (1989).
93. Giraldez, A.J. et al. MicroRNAs regulate brain morphogenesis in zebrafish. *Science* **308**, 833-838 (2005).
94. Xiao, C. et al. MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb. *Cell* **131**, 146-159 (2007).
95. Hornstein, E. & Shomron, N. Canalization of development by microRNAs. *Nat Genet* **38 Suppl**, S20-24 (2006).
96. Wu, C.I., Shen, Y. & Tang, T. Evolution under canalization and the dual roles of microRNAs: a hypothesis. *Genome Res* **19**, 734-743 (2009).
97. Peterson, K.J., Dietrich, M.R. & McPeek, M.A. MicroRNAs and metazoan macroevolution: insights into canalization, complexity, and the Cambrian explosion. *Bioessays* **31**, 736-747 (2009).
98. Girard, A., Sachidanandam, R., Hannon, G.J. & Carmell, M.A. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* **442**, 199-202 (2006).
99. Aravin, A. et al. A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442**, 203-207 (2006).
100. Grivna, S.T., Beyret, E., Wang, Z. & Lin, H. A novel class of small RNAs in mouse spermatogenic cells. *Genes Dev* **20**, 1709-1714 (2006).
101. Vagin, V.V. et al. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**, 320-324 (2006).
102. Saito, K. et al. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the Drosophila genome. *Genes Dev* **20**, 2214-2222 (2006).
103. Houwing, S. et al. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* **129**, 69-82 (2007).
104. Murchison, E.P. et al. Conservation of small RNA pathways in platypus. *Genome Res* **18**, 995-1004 (2008).
105. Lau, N.C., Ohsumi, T., Borowsky, M., Kingston, R.E. & Blower, M.D. Systematic and single cell analysis of Xenopus Piwi-interacting RNAs and Xiwi. *EMBO J* **28**, 2945-2958 (2009).
106. Lau, N.C. et al. Characterization of the piRNA complex from rat testes. *Science* **313**, 363-367 (2006).
107. Li, C. et al. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* **137**, 509-521 (2009).
108. Malone, C.D. et al. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. *Cell* **137**, 522-535 (2009).
109. Palakodeti, D., Smielewska, M., Lu, Y.C., Yeo, G.W. & Graveley, B.R. The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians. *RNA* **14**, 1174-1186 (2008).
110. Friedlander, M.R. et al. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26**, 407-415 (2008).
111. Carmell, M.A. et al. MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev Cell* **12**, 503-514 (2007).

112. Chen, Y., Pane, A. & Schupbach, T. Cutoff and aubergine mutations result in retrotransposon upregulation and checkpoint activation in Drosophila. *Curr Biol* **17**, 637-642 (2007).
113. Cox, D.N. et al. A novel class of evolutionarily conserved genes defined by piwi are essential for stem cell self-renewal. *Genes Dev* **12**, 3715-3727 (1998).
114. Cox, D.N., Chao, A. & Lin, H. piwi encodes a nucleoplasmic factor whose activity modulates the number and division rate of germline stem cells. *Development* **127**, 503-514 (2000).
115. Deng, W. & Lin, H. miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. *Dev Cell* **2**, 819-830 (2002).
116. Kuramochi-Miyagawa, S. et al. Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development* **131**, 839-849 (2004).
117. Brennecke, J. et al. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089-1103 (2007).
118. Gunawardane, L.S. et al. A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in Drosophila. *Science* **315**, 1587-1590 (2007).
119. Aravin, A.A. et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol Cell* **31**, 785-799 (2008).
120. Pelisson, A. et al. Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. *EMBO J* **13**, 4401-4411 (1994).
121. Kim, A. et al. Retroviruses in invertebrates: the gypsy retrotransposon is apparently an infectious retrovirus of Drosophila melanogaster. *Proc Natl Acad Sci U S A* **91**, 1285-1289 (1994).
122. Lecher, P., Bucheton, A. & Pelisson, A. Expression of the Drosophila retrovirus gypsy as ultrastructurally detectable particles in the ovaries of flies carrying a permissive flamenco allele. *J Gen Virol* **78 (Pt 9)**, 2379-2388 (1997).
123. Song, S.U., Kurkulos, M., Boeke, J.D. & Corces, V.G. Infection of the germ line by retroviral particles produced in the follicle cells: a possible mechanism for the mobilization of the gypsy retroelement of Drosophila. *Development* **124**, 2789-2798 (1997).
124. Lane, N. et al. Resistance of IAPs to methylation reprogramming may provide a mechanism for epigenetic inheritance in the mouse. *Genesis* **35**, 88-93 (2003).
125. de Rooij, D.G. & Grootegoed, J.A. Spermatogonial stem cells. *Curr Opin Cell Biol* **10**, 694-701 (1998).
126. Klattenhoff, C. et al. Drosophila rasiRNA pathway mutations disrupt embryonic axis specification through activation of an ATR/Chk2 DNA damage response. *Dev Cell* **12**, 45-55 (2007).
127. Aravin, A.A., Hannon, G.J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318**, 761-764 (2007).
128. Grivna, S.T., Pyhtila, B. & Lin, H. MIWI associates with translational machinery and PIWI-interacting RNAs (piRNAs) in regulating spermatogenesis. *Proc Natl Acad Sci U S A* **103**, 13415-13420 (2006).
129. Wang, G. & Reinke, V. A C. elegans Piwi, PRG-1, regulates 21U-RNAs during spermatogenesis. *Curr Biol* **18**, 861-867 (2008).
130. Das, P.P. 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**, 79-90 (2008).
131. Batista, P.J. et al. PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Mol Cell* **31**, 67-78 (2008).
132. Shendure, J. & Ji, H. Next-generation DNA sequencing. *Nat Biotechnol* **26**, 1135-1145 (2008).
133. Wheeler, D.A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872-876 (2008).
134. Pomraning, K.R., Smith, K.M. & Freitag, M. Genome-wide high throughput analysis of DNA methylation in eukaryotes. *Methods* **47**, 142-150 (2009).
135. Nagalakshmi, U. et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**, 1344-1349 (2008).
136. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* **10**, 57-63 (2009).
137. Park, P.J. ChIP-seq: advantages and challenges of a maturing technology. *Nat Rev Genet* **10**, 669-680 (2009).
138. Berezikov, E. et al. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38**, 1375-1377 (2006).

139. Friedlander, M.R. et al. High-resolution profiling and discovery of planarian small RNAs. *Proc Natl Acad Sci U S A* **106**, 11546-11551 (2009).
140. Stoeckius, M. et al. Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. *Nat Methods* **6**, 745-751 (2009).
141. Shi, W., Hendrix, D., Levine, M. & Haley, B. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat Struct Mol Biol* **16**, 183-189 (2009).
142. Umbach, J.L., Nagel, M.A., Cohrs, R.J., Gilden, D.H. & Cullen, B.R. Analysis of human alphaherpesvirus microRNA expression in latently infected human trigeminal ganglia. *J Virol* **83**, 10677-10683 (2009).
143. Pant, B.D. et al. Identification of nutrient-responsive Arabidopsis and rapeseed microRNAs by comprehensive real-time polymerase chain reaction profiling and small RNA sequencing. *Plant Physiol* **150**, 1541-1555 (2009).
144. Kato, M., de Lencastre, A., Pincus, Z. & Slack, F.J. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol* **10**, R54 (2009).
145. Soares, A.R. et al. Parallel DNA pyrosequencing unveils new zebrafish microRNAs. *BMC Genomics* **10**, 195 (2009).
146. Moxon, S. et al. A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics* **24**, 2252-2253 (2008).
147. Hackenberg, M., Sturm, M., Langenberger, D., Falcon-Perez, J.M. & Aransay, A.M. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* **37**, W68-76 (2009).
148. Mendes, N.D., Freitas, A.T. & Sagot, M.F. Current tools for the identification of miRNA genes and their targets. *Nucleic Acids Res* **37**, 2419-2433 (2009).
149. Bentwich, I. Prediction and validation of microRNAs and their targets. *FEBS Lett* **579**, 5904-5910 (2005).
150. Lindblad-Toh, K. et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803-819 (2005).
151. Zhou, B., Wang, S., Mayr, C., Bartel, D.P. & Lodish, H.F. miR-150, a microRNA expressed in mature B and T cells, blocks early B cell development when expressed prematurely. *Proc Natl Acad Sci U S A* **104**, 7080-7085 (2007).
152. He, L. et al. A microRNA polycistron as a potential human oncogene. *Nature* **435**, 828-833 (2005).
153. Eis, P.S. et al. Accumulation of miR-155 and BIC RNA in human B cell lymphomas. *Proc Natl Acad Sci U S A* **102**, 3627-3632 (2005).
154. Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in C. elegans. *Curr Biol* **13**, 807-818 (2003).
155. Ohler, U., Yekta, S., Lim, L.P., Bartel, D.P. & Burge, C.B. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. *RNA* **10**, 1309-1322 (2004).
156. Reddien, P.W. & Sanchez Alvarado, A. Fundamentals of planarian regeneration. *Annu Rev Cell Dev Biol* **20**, 725-757 (2004).
157. Han, J. et al. Posttranscriptional crossregulation between Drosha and DGCR8. *Cell* **136**, 75-84 (2009).
158. Kadener, S. et al. Genome-wide identification of targets of the drosha-pasha/DGCR8 complex. *RNA* **15**, 537-545 (2009).
159. Linsen, S.E. et al. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* **6**, 474-476 (2009).
160. O'Donnell, K.A. & Boeke, J.D. Mighty Piwis defend the germline against genome intruders. *Cell* **129**, 37-44 (2007).
161. Reddien, P.W., Oviedo, N.J., Jennings, J.R., Jenkin, J.C. & Sanchez Alvarado, A. SMEDWI-2 is a PIWI-like protein that regulates planarian stem cells. *Science* **310**, 1327-1330 (2005).
162. Kawaji, H. et al. Hidden layers of human small RNAs. *BMC Genomics* **9**, 157 (2008).
163. Metropolis, N. & Ulam, S. The Monte Carlo method. *J Am Stat Assoc* **44**, 335-341 (1949).
164. Mockler, T.C. et al. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* **85**, 1-15 (2005).
165. Chen, K. & Rajewsky, N. The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet* **8**, 93-103 (2007).