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

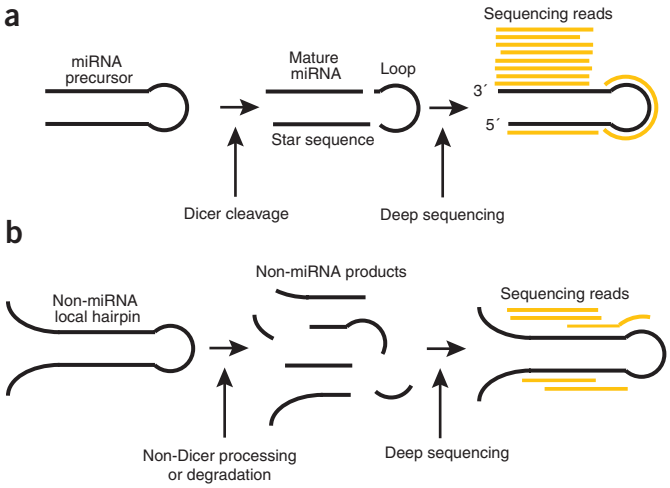

Figure 1 Analyzing the compatibility of sequenced RNAs with miRNA biogenesis. (**a**) Each of the RNA products generated after a stable miRNA precursor is cleaved by Dicer—the mature miRNA sequence, the star sequence and the loop[2]—has a certain probability of being sequenced. When miRDeep maps the 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

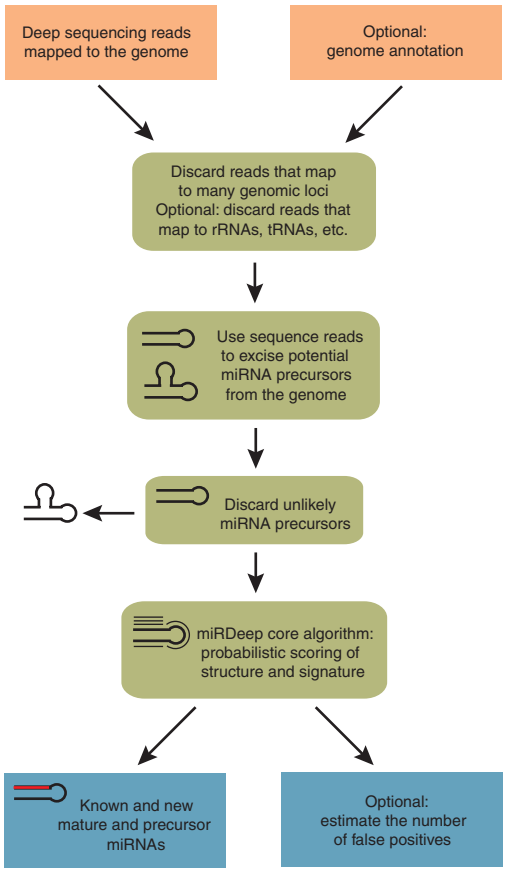

Figure 2 Flowchart diagram representing the miRDeep software package.

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**).

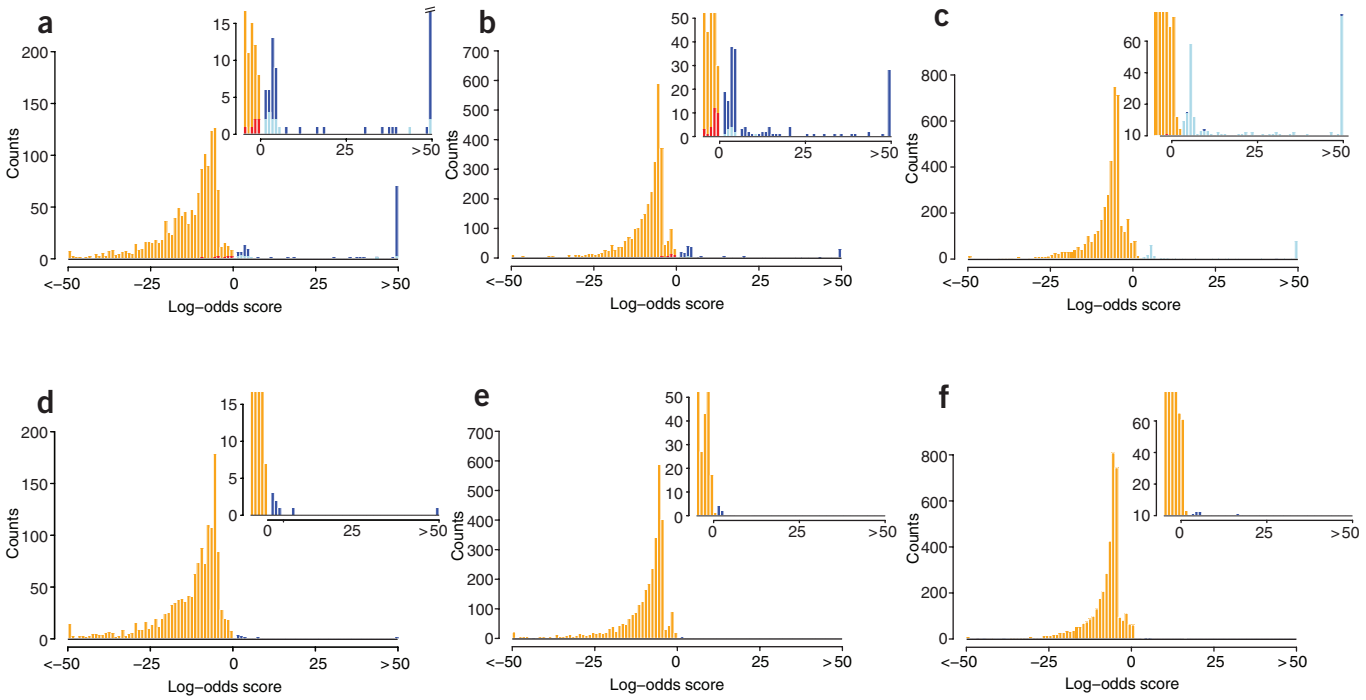

Figure 3 Discovery of known and novel miRNAs by miRDeep. (**a**–**c**) Histograms of miRDeep scores are shown for *C. elegans* (**a**), human (**b**) and dog (**c**) data. The inserts are close-ups. Known miRNA precursors are colored dark blue. False negatives (known miRNA precursors that do not exceed 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

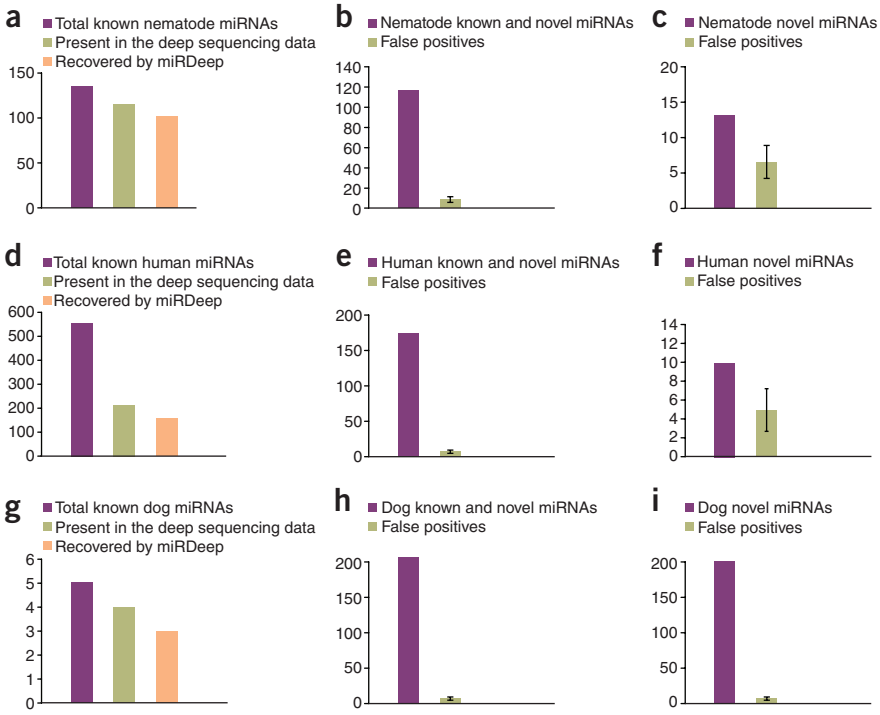

Figure 4 Accuracy of the miRDeep algorithm. (**a**–**i**) The three rows show sensitivity and false-positive estimates for miRDeep results from *C. elegans* (**a**–**c**), human (**d**–**f**) and dog (**g**–**i**). In **a**, **d** and **g**, the total number of mature miRNA sequences known in each species 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.

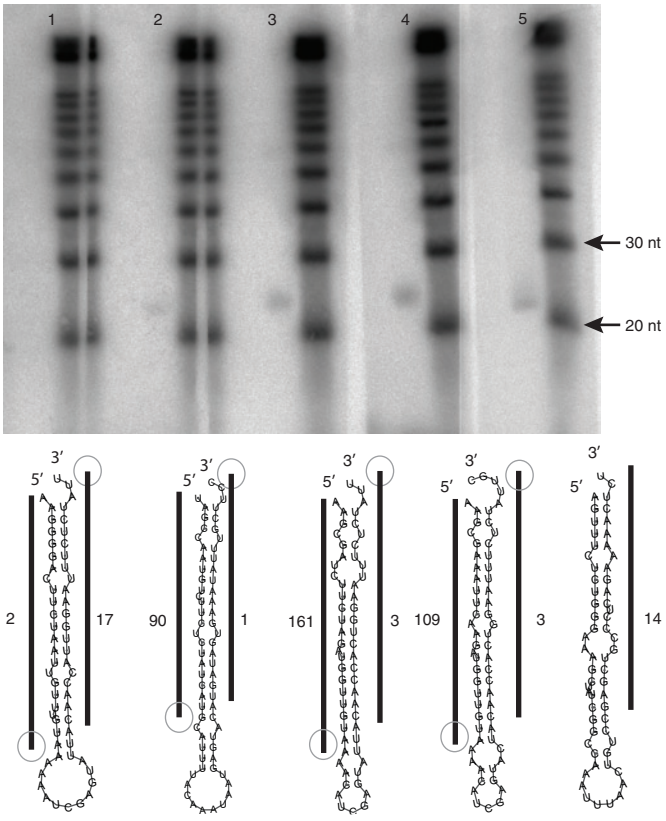

Figure 5 Validating miRDeep candidates by northern blot analysis. Northern blot analysis of five of the novel *C. elegans* miRDeep miRNAs revealed bands corresponding to the mature miRNA product in four out of five candidates (lanes 2–5). The nucleotide length of the mature products as indicated by the 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).