Published online 12 September 2011

Nucleic Acids Research, 2012, Vol. 40, No. 1 ***37–52***
doi:10.1093/nar/gkr688

miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades

Marc R. Friedländer[1], Sebastian D. Mackowiak[1], Na Li[2], Wei Chen[2] and Nikolaus Rajewsky[1,*]

[1]Laboratory for Systems Biology of Gene Regulatory Elements and [2]Laboratory for New Sequencing Technology, Berlin Institute for Medical Systems Biology at the Max-Delbrück-Center for Molecular Medicine, Berlin-Buch 13125, Germany

Received April 20, 2011; Revised August 2, 2011; Accepted August 7, 2011

ABSTRACT

microRNAs (miRNAs) are a large class of small non-coding RNAs which post-transcriptionally regulate the expression of a large fraction of all animal genes and are important in a wide range of biological processes. Recent advances in high-throughput sequencing allow miRNA detection at unprecedented sensitivity, but the computational task of accurately identifying the miRNAs in the background of sequenced RNAs remains challenging. For this purpose, we have designed miRDeep2, a substantially improved algorithm which identifies canonical and non-canonical miRNAs such as those derived from transposable elements and informs on high-confidence candidates that are detected in multiple independent samples. Analyzing data from seven animal species representing the major animal clades, miRDeep2 identified miRNAs with an accuracy of 98.6–99.9% and reported hundreds of novel miRNAs. To test the accuracy of miRDeep2, we knocked down the miRNA biogenesis pathway in a human cell line and sequenced small RNAs before and after. The vast majority of the >100 novel miRNAs expressed in this cell line were indeed specifically downregulated, validating most miRDeep2 predictions. Last, a new miRNA expression profiling routine, low time and memory usage and user-friendly interactive graphic output can make miRDeep2 useful to a wide range of researchers.

INTRODUCTION

microRNAs (miRNAs) are small non-coding RNAs that post-transcriptionally regulate the expression of target mRNAs. The majority of animal miRNAs are transcribed as long primary transcripts from which one or more ~70 nt long hairpin precursors (pre-miRNAs) are cleaved out by the Drosha endonuclease (1). The pre-miRNAs are exported to the cytosol where they are cleaved by the Dicer protein, releasing the loop of the hairpin and a ~22 nt duplex consisting of the mature miRNA and the star miRNA. The duplex is unwound and the mature miRNA is incorporated into the miRNA-induced silencing complex (miRISC) which it can guide to target sites in the 3′ UTRs of mRNA transcripts. This effector complex then either reduces the stability of the mRNA or inhibits its translation (2). Since it is estimated that the transcripts of between 30% and 60% of all human protein coding genes are targeted by one or more miRNAs in one or more cellular contexts (3,4) it is not surprising that miRNAs are involved in almost all biological processes, ranging from development to metabolic regulation and cancer (5–7).

miRNAs must be detected and annotated before their biological functions can be unraveled. While the first miRNAs were detected by conventional cloning and Sanger sequencing (8–10), recent advances in high-throughput sequencing has allowed detection of more lowly abundant miRNAs with unprecedented sensitivity. The algorithms that mine the high-throughput sequencing data for miRNAs use the same basic principles as the algorithms first used to mine the Sanger data, specifically the presence of multiple sequenced RNAs corresponding to the mature miRNA and the presence of a hairpin structure. If the star miRNA or loop is also sequenced this counts as additional evidence. However, the miRNAs detected by the high-throughput platforms are often as lowly abundant as sequenced degradation products of annotated or un-annotated transcripts, making classification much more difficult. Therefore algorithms that mine high-throughput data use advanced post-filtering steps in

* To whom correspondence should be addressed. Tel: +49 30 9406 2999; Fax: +49 30 9406 3068; Email: rajewsky@mdc-berlin.de
Present address:
Marc R. Friedländer, Centre for Genomic Regulation, Barcelona 08003, Catalonia, Spain

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

addition to the basic principles. The miRDeep algorithm, developed by our own lab, uses Bayesian statistics to score the fit of sequenced RNAs to the biological model of miRNA biogenesis (11). MIReNA uses combinatorial rules to identify miRNAs (12). miRanalyzer uses a support vector machine (SVM) trained on miRNA features to classify miRNA transcripts from non-miRNA transcripts (13,14). miRTRAP identifies gene loci where many sequenced RNAs map to few defined positions (15). Evaluation of these algorithms is however difficult since they have each only been tested on a limited number of data sets representing limited coverage of the animal phylogenetic tree. Furthermore, validation of the reported novel miRNAs has either been restricted to few candidates (miRDeep, miRTRAP) or not performed (miRanalyzer). To address this problem of evaluation, we propose that a method to identify miRNAs in high-throughput sequencing data should meet three demands. Specifically we demand that the method:

(1) can accurately identify known and novel miRNAs in all animal major clades;
(2) can distinguish miRNAs from other argonaute-bound small RNAs;
(3) reports miRNAs that can stand up to high-throughput validation.

Besides the method should ideally:

(4) be efficient in memory and time consumption;
(5) be user-friendly.

To meet these demands, we have completely overhauled our original miRDeep algorithm and added extensive new packages. In this article, we describe these changes and extensions. miRDeep2 has internal statistical controls that allow to estimate the accuracy and sensitivity of its performance. To test miRDeep2 performance by an independent method, we present experiments in which we knocked down the miRNA pathway and monitored changes in expression of known miRNAs, novel miRDeep2 miRNAs and other small RNA classes.

MATERIALS AND METHODS

miRDeep2 module

This section describes the default work-flow of the miRDeep2 module in detail. The first step tests the format of input files (see online documentation for format requirements).

After that a fast quantification of known miRNAs is done if files with miRBase precursors and corresponding mature miRNAs are given to the module. In a second step, potential miRNA precursors are excised from the genome using the read mappings as guidelines. The read mappings are first parsed such that only perfect mappings (no mismatches) of at least 18 nt are retained. Furthermore, read mappings from reads that map perfectly more than five times to the genome are discarded. Then the two genome strands of each genome contig are scanned separately, from 5′ to 3′ end. Excision is initiated when a stack of reads (height one or more) is encountered. If there is a higher read stack within 70 nt downstream of the current read stack, then this is chosen instead. This downstream search is iterated until no higher read stack is found within 70 nt. In this way, the highest local read stack is identified. Then the sequence covered by the highest local read stack is excised twice, once including 70 nt upstream and 20 nt downstream flanking sequence, and once including 20 nt upstream and 70 nt downstream flanking sequence. Subsequently, the genome scanning continues from the position 1 nt downstream of the last excised sequence. If the total number of potential precursor sequences excised is less than 50 000 (two precursors per genomic locus), then this set is output to the downstream analysis. If there are more sequences, then the entire excision step is repeated, with the height of the read stack necessary for initiating excision increased by one. The third step of the module is to prepare the signature file. The bowtie-build tool is used with default options to build a Burrows-Wheeler transform index of the excised potential precursors. Then the set of sequencing reads is mapped to the index, using bowtie (version 0.12.7) with the following options: bowtie –f –v 1 –a –best –strata –norc. Option '-f' designates a fasta file as input, option '-v 1' reports read mappings with up to one mismatch to the precursors, option '-a' leads to the report of all valid alignments, options '–best –strata' orders the mappings from best to worse alignments according to the strata definition of bowtie. If reads map perfectly to the precursors then mappings of the same read with one mismatch are not reported. Option '–norc' advises bowtie not to map reads to the reverse complement of the precursor sequences in the bowtie index. The set of known mature miRNAs for the reference species is also mapped to the index, with the following options: bowtie –f –v 0 –a –best –strata –norc. Here we do not allow any mismatches for the mappings because the mature miRNA sequence and the potential precursor sequences have not been subject to any source of noise.

The two mapping files are concatenated and all lines are sorted according to the potential precursor ids. The fourth step is to predict RNA secondary structures of the potential precursors. This is done with RNAfold with default options. Optionally, the randfold *P*-values for a subset of the potential precursors are calculated. This is done by selecting the potential precursors that (i) fold into an unbifurcated hairpin, (ii) can be partitioned into candidate mature, loop and star part based on the reads mapping to it, (iii) have minimum 60% of the nucleotides in the candidate mature part base paired. The randfold *P*-values are calculated for the subset of potential precursors with these options: randfold –s 99. In the fifth step the potential precursors are individually scored or discarded by the miRDeep2 core algorithm. The core algorithm is identical to the first version (11), except for: (i) all mappings to the anti-sense strands of potential precursors are ignored (ii) potential precursors are discarded if <60% of the nucleotides in the candidate mature part are base paired. This displaces the rule that potential precursors are discarded if <14 nt in the candidate mature part are base paired. The miRDeep2 core algorithm is run with these

options: -s –v -50 –y. Option '-s' designates the reference mature miRNAs file in fasta format as input, option '-v -50' keeps all precursors that have a miRDeep2 score above −50 and option '-y' supplies an additional file with randfold values. Furthermore, 100 rounds of permuted controls are performed as previously described (11), with same options as the genuine run. The sixth step surveys the score distributions of the genuine run and the control runs. The performance statistics are calculated for all score cut-offs from −10 to 10. The number of known miRNAs present in the data is estimated as the number of known mature miRNAs that map perfectly to one or more excised potential precursors. The number of known miRNAs that are recovered is estimated as the number of known mature miRNAs that map perfectly to one or more hairpins that exceed the given score cut-off. The sensitivity of the run is estimated as se = (known miRNAs recovered)/(known miRNAs in data). The number of false positives for a given score cut-off is estimated by the permuted controls. The fraction of true miRNAs reported is estimated by t = (novel miRNAs – estimated false positive novel miRNAs)/novel miRNAs. The signal-to-noise ratio is estimated as n = total miRNAs/estimated total false positive novel miRNAs (total miRNAs = novel miRNAs + known miRNAs).

Mapper module

This section describes the default work-flow of the Mapper module in detail. The first step tests the format of the input files. The second step parses the raw Solexa/Illumina _seq.txt output file into fasta format. Raw solexa output files are text files that contain one line per sequenced read. These can be parsed and transformed to other text file formats like fastq/fasta files. The third step clips 3′ adapters and collapses reads. The read sequence is searched for matches to the first 6 nt in the adapter sequence. This search starts at position 18 in the read. If there are no matches to the first 6 nt, then matches to the first 5 nt of the adapter are searched in the last five nt of the read, then matches of the first fours to the last four positions and so on. When a match is first found, the match to the adapter sequence and all nucleotides downstream are clipped from the read, and the next read is searched. Reads that have no matches are retained, but not clipped. Next, all reads with identical sequence are collapsed to remove redundancy. A digit in the new fasta identifiers shows how many times the corresponding sequence was present in the data set. The fifth step maps the processed reads to the genome with bowtie, using these options: bowtie –f –n 0 –e 80 –l 18 –a –m 5 –best –strata. Option '-n 0' keeps only alignments with 0 mismatches in the seed region of a read mapped to the genome. The seed region is defined by option '-l 18' that corresponds to the first 18 nt of a read sequence. When using option '-n' it is possible to allow mismatches occurring after the seed region of a read in an alignment. This is determined by option '-e 80' and is the maximum sum of quality values at each mismatch position. The default quality value for each position in a fasta file is set to 40 which means that up to two mismatches are allowed in the region of a read after its seed region. Option '-m 5' keeps only reads that do not map more than five times to the genome. Option '–best –strata' orders the mappings from best to worse alignments according to the strata definition of bowtie. If mappings with zero mismatches occur then mappings with one or two mismatches are not reported. Finally the processed reads and the mappings to the genome are outputted. Other mapping tools such as BWA (16) can be used, but their output format needs to be converted into the .arf format (see online documentation). However, the next miRDeep2 update will support BWA as a mapping tool.

Quantifier module

This section describes the default work-flow of the Quantifier module in detail. The first step tests the format of the input files. The second step maps the sequencing reads, the known mature miRNAs and optionally its star sequences for the reference species against the known precursor miRNAs for the reference species. The bowtie-build tool is used with default options to build a Burrows-Wheeler transform index of the known precursors. The mapping of the reads is done with these options: bowtie –f –v 1 –a –best –strata –norc. Option '-f' designates a fasta file as input, option '-v 1' reports read mappings with up to one mismatch, option '-a' leads to the report of all valid alignments, options '–best –strata' orders the mappings from best to worse alignments according to the strata definition of bowtie and option '–norc' advises bowtie not to map reads to the reverse complement of the precursor sequences in the bowtie index. The mapping of the known mature and star miRNA sequences against the known precursor miRNAs for the reference species is done with these options: bowtie –f –v 0 –a –best –strata –norc. Here the number of allowed mismatches is set to zero via option '-v' because annotated mature and star sequences should be contained in their annotated precursor sequences. Mappings of mature miRNAs to unmatched precursors are discarded (for instance, the only miR-9 mapping that is retained is to the mir-9 precursor). The third step intersects the two mapping files. A read is assumed to represent a sequenced mature miRNA if it falls within the same position on the precursor, plus 2 nt upstream and 5 nt downstream. We allow a small window around the annotated mature miRNA in its precursor because reads originating from real miRNAs can be subject to untemplated nucleotide addition and unprecise Dicer processing. Reads that map equally well to the positions of two or more mature miRNAs are added to the read counts of all of those mature miRNAs.

miRDeep2 analysis of sequenced small RNAs from seven animal species

The following sequencing data set series were downloaded from the GEO database (17): human liver, GSE21279 (only the three healthy human liver samples were analyzed); human cell lines, GSE16579 (only the human data from this series were analyzed); mouse, GSE20384 (only the mouse data from this series were analyzed); sea squirt, GSE21078, GSE13625; fruit fly, GSE7448;

nematode, GSE17153; planarians, GSE16159 (Illumina data only); sea anemone, GSE12578 (from this series, only the sea anemone data were analyzed). Furthermore, the following data sets were retrieved from the SRA database (18): nematode, SRR014966-73. For each of the seven species, all of the available data was pooled (files concatenated) at the first point in the analysis when all the data had the same format. The raw Illumina _seq.txt data sets were processed and mapped against the reference genomes with the Mapper module using these options: -a –h –i –j –k TCGTAT –l 18 -m –p –s –t. Option '-a' designates that the reads file is in _seq.txt format, option '-h' converts the raw reads file to fasta format, option '-i' converts RNA to DNA alphabet in reads file, option '-j' removes all read-sequences that contain letters other then A, C, G, T, U, N, option '-k' specifies the sequencing adapter for clipping reads, option '-l' designates minimum read length after adapter clipping, option '-m' collapses reads, option '-p' designates the bowtie index to which reads are mapped, option '-s' specifies the processed reads file name and option '-t' is the name of the mapping file that contains the read mappings. The GEO tabular files and the SRA fastq files were parsed into fasta format with custom perl scripts and processed and mapped against the reference genomes using the Mapper module with these options: -c –i –j –l 18 -m –p –s –t. Option '-c' designates that the reads file is already in fasta format. For each species, miRDeep2 was run on the data with default options, taking the following input files: reads in fasta format, genome in fasta format, read mappings in .arf format, mature and precursor miRNAs from the reference species in fasta format and mature miRNAs from related species in fasta format. Known miRNAs input were all from miRBase version 16. For the miRDeep2 analysis of the human data, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla* and *Pongo pygmaeus* were designated as related species for the purpose of input miRNAs. For the mouse analysis, *Rattus norvegicus* and *Homo sapiens* were considered related species. For the sea squirt analysis, *Danio rerio, Xenopus tropicalis, Mus musculus* and *H. sapiens* were considered. For the fruit fly analysis, *Anopheles gambiae*, *Anopheles mellifera*, *Bombyx mori*, *Locusta migratoria*, *Triboleum castaneum* and all *Drosophila* species were considered related species. For the nematode analysis, *Caenorhabditis briggsae* and *D. melanogaster* were considered related species. For the planarian analysis, *Hydnum rufescens*, *Schistosoma japonicum*, *Schistosoma mansoni*, *D. melanogaster* and *C. elegans* were considered related species. For the sea anemone analysis, *Amphimedon queenslandica* was considered a related species. Note that the species chosen were not in all cases genuinely closely related species. In some cases relatively distant species were chosen because no genuinely closely related species are represented in miRBase (e.g. sea anemone), in other cases relatively distant species were chosen because they have been have been carefully annotated in terms of miRNA genes (e.g. human was designated as related species for mouse). These genome versions were used: human, NCBI hsa v36.3; mouse, UCSC mm9; sea squirt, JGI version 1.0; fruit fly, flybase Dme r5.19; nematode, wormbase ws205; planaria, Smed assembly v31; sea anemone, Nemve1. For each analysis, the lowest score cut-off that yielded a signal-to-noise ratio of 10:1 or higher was used, except for the human liver and sea squirt analyses, where a signal-to-noise ratio of 5:1, respectively, 3.5:1 was used. For a number of species the set of reported miRNA precursors show substantial sequence redundancy: sea squirt, fruit fly, planaria and sea anemone. To be conservative, within each of these sets we identified all precursors that have ≥90% identity over 40 or more nucleotides. These precursors were discarded from our analysis and are not reported in Figure 3. Sensitivity was calculated as *se* = number of miRNA loci above score cut-off/total number of miRNA loci. miRNA loci were defined as the genome positions to which known mature miRNAs present in the data map, counting multiple miRNA loci with identical mature sequences only once. Known miRNAs were defined as all miRBase version 16 mature miRNAs for the species analyzed. Note that this definition of sensitivity is equivalent to the definition in the 'Materials and Methods' section on the miRDeep2 module. Specificity was calculated in the following way. Precursors were excised from the genome using the mapped reads as guidelines as described in the 'Materials and Methods' section on the miRDeep2 module. The precursors that did not have any known miRNAs mapping were for the purposes of this analysis considered non-miRNA loci. The specificity was then calculated as *sp* = number of non-miRNA loci below the score cut-off/total number of non-miRNA loci. The prevalence was calculated as p = known miRNA loci/total number of precursors excised. Last, accuracy was calculated as: $a = se \times p + sp(1 - p)$. The true positive rate was calculated as described in the 'Materials and Methods' section on the miRDeep2 module. The positive predictive value was calculated as: $ppv = tp/(tp + fp)$ where tp = known miRNAs in data × sensitivity + novel miRNAs reported × estimated true positive rate and fp = novel miRNAs reported × (1–estimated true positive rate). Novel miRNAs were considered high-confidence if both the putative mature and star miRNA reported by miRDeep2 were detected in at least two independent samples, having the exact same 5′- and 3′-ends and allowing no mismatches.

Dicer silencing, qPCR measurements and sequencing of small RNAs in MCF-7 cells

A total of 20 nM of siDicer duplex was transfected into MCF-7 cells with Lipofectamine RNAiMax transfection reagent (Invitrogen) according to the manufacturer's instructions. Target sequence of siDicer is 5′-UGCUUGAA GCAGCUCUGGA-3′. Nuclear and cytoplasmic extracts were prepared using PARIS kit (Ambion). Briefly, to prepare cytoplasmic extracts, cells were harvested by trypsinization, washed and pelleted by centrifugation. The pellet was resuspended in 1× PBS and pelleted again by spinning. After removing PBS, the cell pellet was resuspended in ice-cold cell fractionation buffer by gentle pipetting. The homogenate was centrifuged, and the supernatant containing the cytoplasm was transferred to a fresh

tube and subsequently used for RNA extraction and small RNA sequencing. The remained pellet was washed with ice-cold cell fractionation buffer and designated the nucleus. Total RNA from cell pellets was extracted using TRIZOL reagent (Invitrogen) and total RNA from cytoplasm was isolated using TRIZOL LS reagent (Invitrogen) following the manufacturer's protocol, respectively. Total RNA was treated with DNase I (Ambion). Small RNA fraction with size range of 10–40 nt was separated from total RNA using flashPAGE Fractionator (Ambion) according to the manufacturer's instruction. mRNA and miRNA RT-quantitative PCR studies were carried out using SYBER Green assay and Taqman assay systems (Applied Biosystems), respectively. mRNA expression was normalized to GAPDH and miRNA expression was normalized to RNU48. Small RNA libraries were prepared for Illumina deep-sequencing. Briefly, the small RNA fraction was ligated sequentially at the 3′OH and 5′phosphates with synthetic RNA adapters, reverse transcribed and amplified using Illumina sequencing primers. Finally, the adapter-ligated libraries were sequenced for 36 cycles on the Illumina/GA II platform, according to the manufacturer's instructions.

Calculation of small RNA expression fold-changes upon *Dicer* silencing

The C/D box, H/ACA box and Cajal-body-specific snoRNA sequences were obtained from snoRNABase at http://www-snorna.biotoul.fr. The tRNA sequences were obtained from tRNAdb at http://trnadb.bioinf.uni-leipzig.de/ (19). The four consensus rRNA sequences (5S, 5.8S, 18S and 28S) were obtained from NCBI at http://www.ncbi.nlm.nih.gov/. The initial set of 26 241 control sequences consisted of all potential precursors excised by miRDeep2 but discarded before being assigned a miRDeep2 score. The set of 940 known human miRNA precursors were downloaded from miRBase version 16 at http://www.mirbase.org/ (20). The miRDeep2 precursors were produced by analyzing the human cell line data with default options and a score cut-off of 4, considering only perfectly mapping reads. The small RNA reads produced by sequencing the four MCF-7 samples were clipped of 3′ adapters using the clip_adapters.pl script from the miRDeep2 package. The two data sets produced by sequencing the unperturbed cells were pooled, as were the two data sets produced by sequencing the cells exposed to *Dicer* silencing. These pooled sets were independently mapped to the following sequences: snoRNAs, tRNAs, rRNAs, control sequences, miRBase precursors and miRDeep2 precursors using bowtie with the following options: -f -v 2 -a –best –strata –norc. For each annotated sequence, the sum of reads mapping from the unperturbed and the *Dicer* silenced sample was calculated. If this sum was <40, the sequence was not considered and is not plotted in Figure 4D–I. If this sum was 40 or higher, the log2 fold-change was calculated as follows: $f = \log2$ (number of reads mapping from *Dicer* silenced sample/number of reads mapping from unperturbed sample). To perform a complementary analysis with genomic control sequences that are independent of the miRDeep2 excision procedure, the following was done. The reads from unperturbed and *Dicer* silenced samples were mapped to the human genome assembly NCBI hsa v36.3 using bowtie with the following options: -f -v 2 -a -m 1 –best –strata. Then the human genome was divided into non-overlapping regions of 100 nt. All the regions that harbored miRBase precursors were assigned as miRBase regions, all regions that harbored miRDeep2 regions were assigned as miRDeep2 regions and all remaining regions were assigned as control regions. Then for each region, the log2 fold-change of reads mapping was calculated as above, only considering regions to which 40 or more reads mapped. This analysis yielded comparable results to the analysis described above (Figure 4G–I), as did a similar analysis where the reads were mapped with options: -f -v 2 -a -m 100 –best –strata, and each read was subsequently weighed inversely to the number of genome mappings. To investigate how miRDeep2 compares with competing methods, we used four programs to analyze the same data, consisting of the human cell line data, considering only reads that map perfectly to the human genome (hg19). miRDeep2 and MIReNA were run with default options to produce 509 and 288 predictions, respectively. For miRDeep2 these are all predictions with a score of 0 or higher. miRTRAP was run with default options and minLocus count of 15 and a minShift of 5 to produce 195 predictions, while miRanalyzer was run with default options and a score cut-off of 1 to yield 1590 predicted precursors. Then we created a set of predictions that is exclusive to each program as well as a set that is common to miRDeep2 and the competing method. These sets were divided into subsets based on the number of reads that support each precursor, summing over the siDicer and control data. Specifically, subsets were created of predictions that are supported by a minimum of 1, 5, 10, 25, 50, 75 or 100 reads. Last, log2 fold-changes were calculated for each subset as above.

Benchmarking on sampled subsets of human small RNAs

From the Gene Expression Omnibus (GEO) and Short Reads Archive (SRA) databases, we compiled human small RNA data from 13 studies, comprising altogether 94 distinct data sets from tissues, cell lines and cancers (21–33). These data were parsed to fasta format and adapters were clipped (where present) and reads <18 nt were removed with the Mapper module using these options: -h -i -j -k -l -m. Then the data from each set were pooled to generate what we refer to in the following as the 'undiluted data set'. The 10 diluted data sets were generated by sampling the undiluted data set separately 10 times. For each sampling, each read in the undiluted data set was retained with a probability equal to the dilution fraction (e.g. to generate the 0.1 dilution data set, each read in the undiluted data set was retained with 10% probability and discarded with 90% probability). Each of the ten data sets was processed and mapped against the human hsa v36.3 genome using the Mapper module with these options: –p –s –t. Then each set of data was analyzed by the miRDeep2 module with default options, taking the following input files: reads in fasta

format, genome in fasta format, read mappings in .arf format, mature miRNAs from the reference species in fasta format and mature miRNAs from related species in fasta format. miRNAs input were all from miRBase version 16, and *P. troglodytes*, *P. paniscus*, *G. gorilla* and *P. pygmaeus* were considered related species. For each analysis, the lowest score cut-off that yielded a signal-to-noise ratio of five or higher was used.

RESULTS

Work flow of miRDeep2 modules

The miRDeep2 package consists of three modules (Figure 1). The miRDeep2 module identifies known and novel miRNAs in high-throughput sequencing data. The Mapper module processes raw sequence output from the Illumina platform and maps the processed reads to the reference genome. The Quantifier module sums up read counts for known miRNAs in a sequencing data set. The modules work complementary, for instance the output of Mapper can be directly input to the miRDeep2 module.

The miRDeep2 module identifies known and novel miRNAs in the analyzed high-throughput sequencing data and forms the core of the software package. The input to miRDeep2 is the reference genome, a set of high-throughput sequencing reads and a file with positions of the reads mapped against the genome (Figure 1A). Optionally, known mature, star and precursor miRNAs from the species analyzed and/or mature miRNAs from related species can be input (see below). The first step of the work flow is to test the format of the input files, so that any format problems are identified and can be corrected by the user before the analysis begins. If known mature and precursor miRNAs for the species analyzed are specified, these are automatically input to the Quantifier module (see below) to ensure that all known miRNAs in

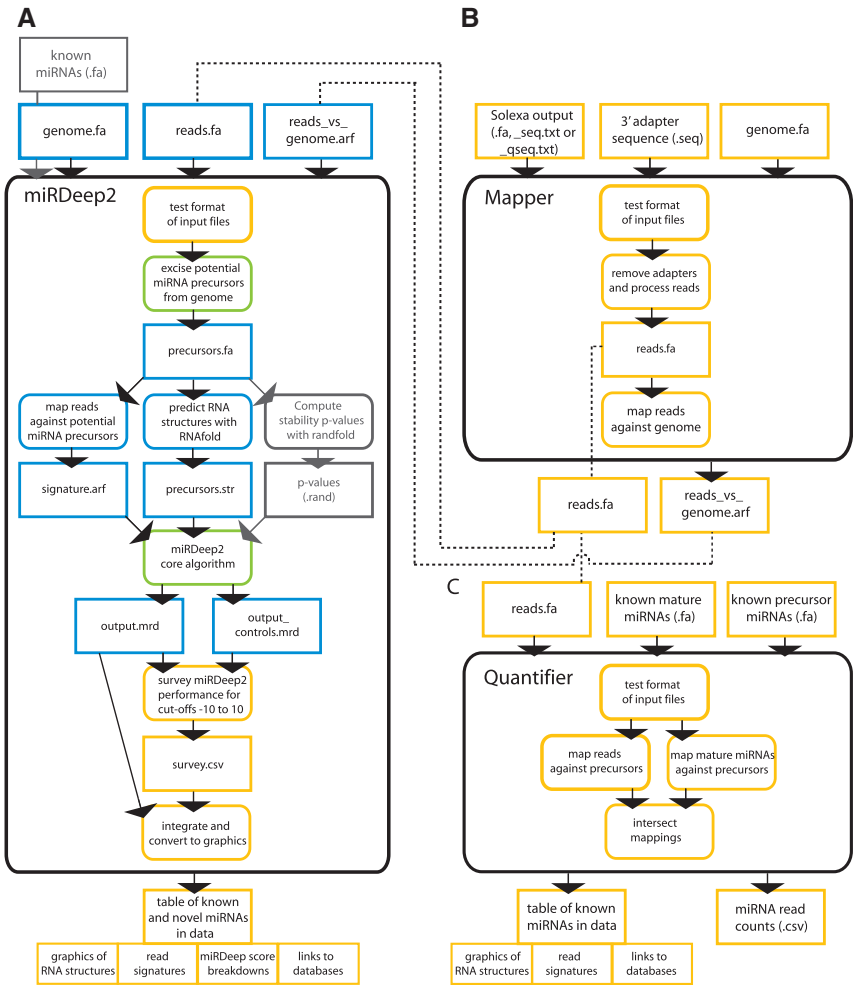

Figure 1. Flow charts of modules. Flow charts for (**A**) the miRDeep2 module (identifies known and novel miRNAs in high-throughput sequencing data), (**B**) the Mapper module (processes Illumina output and maps it to the reference genome) and (**C**) the Quantifier module (sums up read counts for known miRNAs in a sequencing data 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. Files and processes that have been modified are in green. Those that remain largely unchanged from the first version of miRDeep are in blue, while those that are optional are in grey. The file formats are: .fa, fasta; .arf, arf mapping format; .str, RNAfold output; .rand, randfold output; .mrd, miRDeep2 text output; .csv, csv spread-sheet; _seq.txt, raw sequence output from the Illumina platform; seq, sequence given on command line (see online documentation for description of formats). The 'work flow of miRDeep2 modules' results section contains detailed descriptions of all steps.

the data are included in the output table, even if they are not scored by miRDeep2. Second, potential miRNA precursor sequences are excised from the genome, using the mapped reads as guidelines. The algorithm identifies stacks of reads that could be sequenced mature miRNAs and excises the genomic sequence covered by the stack and some flanking sequence (see later section for details). Third, the Bowtie mapping tool (34) is used to map the sequencing reads against the excised miRNA potential precursors (see 'Materials and Methods' section). We refer to the read mappings to a potential precursor as the read 'signature' of the precursor. Fourth, the RNAfold tool (35) is used to predict if the RNA secondary structures of each excised potential precursor resemble a typical miRNA hairpin structure. Since it is known that miRNA precursors are energetically stable given the nucleotide compositions (36), the stability of potential precursors can optionally be predicted using the randfold tool. Fifth, the miRDeep2 core algorithm evaluates the structure and signature of each potential miRNA precursor. If the structure resembles a miRNA hairpin and the reads fall in the hairpin as would be expected from Dicer processing, then the potential precursor is assigned a score that reflects the likelihood of it being a genuine miRNA (11). If not, the potential precursor is discarded. The input to the core algorithm will likely contain a large number of hairpins and a large number of read stacks that have no connection to miRNA biology. The chance intersection between these can produce false positives. The built-in controls of miRDeep2 estimate how large this chance intersection is by shuffling the observed combinations of structures and signatures and re-inputting them to the core algorithm (see 'Materials and Methods' section). The difference in score distributions generated by the controls and the genuine run is then used to estimate the number of true positive novel miRNAs reported by miRDeep2 for varying score cut-offs. The last step of the module integrates all results into an easy to mine overview .html table which contains detailed information on every miRNA identified in the sequencing data (see later section).

The Mapper module is designed as a flexible tool to process and map small RNA sequencing data. The default input is raw text output from the Illumina platform, the 3′ adapter sequence used in the library preparation, and the reference genome (Figure 1B). First, reads undergo processing steps selected by the user. These steps include 3′ adapter removal, length filtering and collapsing of identical read sequences. Then the processed reads are mapped against the reference genome with the Bowtie tool (34). The output of Mapper is a file with the processed sequencing reads and a file with the reads mapped against the genome. These can be input to the miRDeep2 or the Quantifier module or used for other purposes. With default options the genome mapping is stringent compared with the tolerant mapping used by the miRDeep2 module to generate the precursor 'signatures'. We choose the stringent mapping because we do not want to consider loci to which no reads can be traced with high confidence. On the other hand, when generating the read signature of a given locus we are tolerant, since a single miRNA star read with a sequencing error can constitute strong evidence of miRNA biogenesis.

The Quantifier module is designed as a fast and lightweight tool to sum up read counts for known miRNAs in sequencing data. It can be used as a stand-alone tool but is also enacted as part of the miRDeep2 module analysis. The input to the Quantifier module is a set of sequencing reads and known mature and precursor miRNAs from the reference species. Optionally, a file with known star sequences can be given (Figure 1C). First, the module maps the reads and the miRNA strands (mature and star) separately against the precursors. The mappings of the reads and the miRNA strands are then intersected, such that reads that map to the same positions as a given strand add to the read count of that miRNA strand (see 'Materials and Methods' section). The output of Quantifier is a .html table similar to that produced by the miRDeep2 module plus an easy to parse spread-sheet with read counts of all known miRNAs in the data.

Improvements to the miRDeep algorithm

First, miRDeep2 offers a conceptual advance in identifying high-confidence miRNA candidates. In most high-throughput sequencing protocols, small RNA libraries are amplified by PCR reaction before being sequencing. This means that a single small RNA molecule in the sample can give rise to multiple sequencing reads. Thus the fact that a given small RNA is detected multiple times in a given sample does not necessarily constitute evidence that it is prevalent. In contrast, when a small RNA is consistently detected in distinct samples, it does constitute independent evidence that it is prevalent and thus the likely result of a specific biogenesis. According to our definition, two samples are distinct if they underwent amplification in separate PCR tubes. In the beginning of each miRDeep2 analysis, each read is computationally tagged to trace from what sample it originates. During identification of known and novel miRNAs, reads from all samples can be pooled in order to give a more accurate prediction. When the results are reported, the reads are again de-convoluted so the user can see the sample origin of each read. Figure 2 shows a novel human miRNA reported by miRDeep2. Both mature and star strands are detected in three independent liver samples, showing that both strands are prevalent in human liver and thus likely to result from specific biogenesis.

miRDeep2 also offers increased robustness in identifying non-canonical miRNAs that are prevalent in some species. In some invertebrates like sea squirt (37), but also to a lesser extent in mammals (38), particular miRNA precursor hairpins appear to undergo two rounds of Dicer cleavage, resulting in two miRNA duplexes being produced from each hairpin. The miRNAs of the first, non-canonical duplex produced are sometimes referred to as 'moRs' (37). While sequenced miRNAs generally map back to the genome locus in three piles, corresponding to the mature, star and loop sequences, the addition of moRs can thus result in four or five piles. In the first version of miRDeep, excision of potential miRNA hairpins is done by scanning the genome for *clusters* of

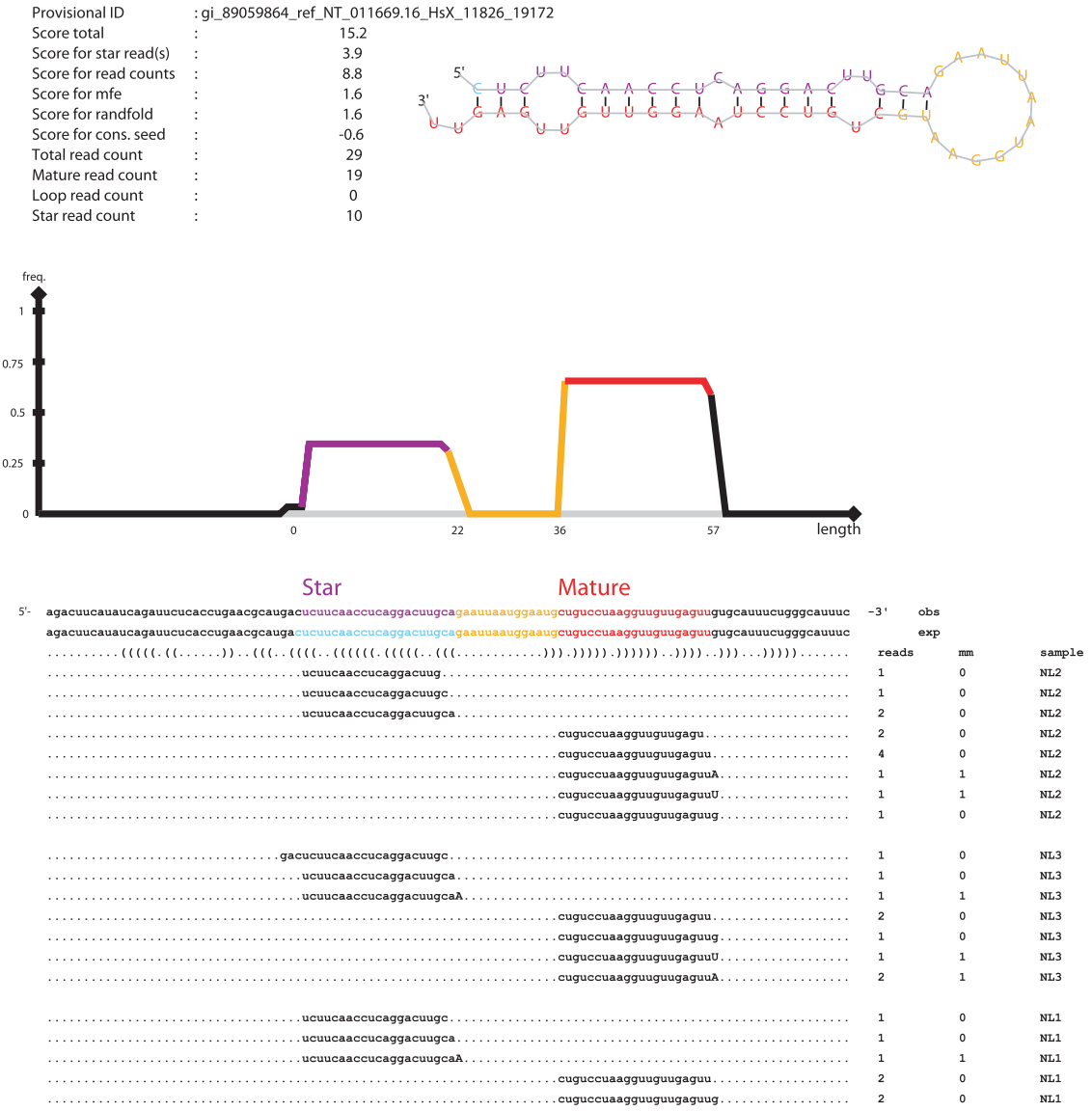

Figure 2. Novel human miRNA detected in three independent liver samples. The upper left table gives the miRDeep2 score break-down for the reported miRNA, along with read counts for the mature, loop and star sequence. The upper right figure shows the predicted RNA secondary structure of the hairpin, partitioned according to miRNA biogenesis: red, mature; yellow, loop; purple, star. The middle density plot shows the distribution of reads in the predicted precursor sequence. The sequences below indicate the positions of the mature, loop and star strand. The positions of the star strand as expected from Drosha/Dicer processing is shown in light blue, while the star consensus positions as observed from the sequencing data is shown in purple. The dotted lines below show the aligned reads. Mismatched nucleotides are presented in upper case. mm, number of mismatches. Both mature and star miRNA strands of this particular miRNA are detected in each of three independent liver samples (NL1-NL3).

mapped reads. We define a cluster as one or more reads that separated from other clusters by minimum 30 nt. The sequence covered by each cluster is excised, including some bracketing sequence (11). In the cases where a single pile of moRs are present, the miRNA precursor is sometimes excised asymmetrically and does not fold into a hairpin structure (not shown). miRDeep2 in contrast performs excision by scanning the genome for *stacks* of reads. We define a stack as one or more reads that map to the exact same 5′ and 3′ positions in the genome. When a stack is encountered, a search identifies the highest local stack (see 'Materials and Methods' section). This stack is assumed to consist of reads from the sequenced mature miRNAs. The sequence covered by the stack is then excised, including some bracketing sequence. Since the excision is based only on the stack of reads from the mature miRNA, the algorithm does not risk an asymmetrical excision due to moRs.

Similarly, there is now solid evidence that functional miRNAs can be transcribed from both genomic strands of a given miRNA locus [e.g. the *D. melanogaster* mir-iab locus, (39)]. The first version of miRDeep considered reads that map anti-sense to a potential miRNA precursor as being inconsistent with Dicer processing, meaning that the reads derived from the genomic minus strand would count against the precursor that is transcribed from the genomic plus strand, and vice versa. In miRDeep2, the two genomic

strands are analyzed separately, meaning that only reads mapping sense to a potential miRNA precursor are considered. miRDeep2 correctly identifies both mir-iab-4 and the mir-iab-8 (anti-sense) miRNAs when analyzing fly data (see later section), giving proof-of-principle that miRDeep2 can detect anti-sense miRNAs.

Last, next generation sequencing reads often contain nucleotides that differ from those of the reference genomic sequence. The reasons for this can be technical, like sequencing errors, or biological, like untemplated nucleotides added to the 3′-end of small RNAs [e.g. (38,40)]. While the first version of miRDeep primarily supported perfect mappings, miRDeep2 supports single or multiple mismatches.

To make a direct comparison of miRDeep2 with the first version of the algorithm, we have run it on the data used in the initial publication, using the exact same input files (genome assemblies, miRBase version etc.) to facilitate the comparison. We find that miRDeep2 performs substantially better on the human data, reporting 186 known and 36 novel miRNAs (compared to 154 known and 10 novel in the initial publication). The improvement is largely due to more accurate detection of lowly abundant miRNAs. We find that miRDeep2 performs slightly better on the nematode data (104 known and 20 novel versus 102 and 13 novel) and about as well as the initial method on the dog data (200 novel and three known versus 203 novel and three known). However, consistent with the implemented changes we find that the most notable improvement in accuracy is in identification of non-canonical miRNAs like the moRs or the anti-sense miRNAs. When identifying miRNAs in data from sea squirts, known to harbor large numbers of non-canonical miRNAs, the first version of miRDeep only reports 46 known and 31 novel miRNAs. In contrast miRDeep2 reports 313 known and 127 novel ones (see next section).

miRDeep2 identifies known and novel miRNAs with high accuracy in seven animal clades

Since the first version of miRDeep was published, sequenced small RNA libraries from numerous model systems have become available from public databases (5,15,25,37,38,41–44). This has allowed us to re-test the claim that miRDeep performs species-independent miRNA discovery by scoring gene features that are shared by animals. We obtained data from seven animal species representing vertebrates, non-vertebrate deuterostomes, ecdysozoans, lophotrochozoans and non-bilaterians (see Figure 3 and 'Materials and Methods' section). With the exception of human and mouse, these species all diverged from each other more than 500 million years ago. For each of the species, all available data were pooled and then processed and mapped to the reference genome with the Mapper module (see 'Materials and Methods' section). Then for each of the seven species the processed reads, mappings, reference genome and miRBase version 16 known miRNAs were input to the miRDeep2 module. The default options were used to analyze data for all species.

The results of the seven runs are shown in Figure 3 (Supplementary Figures S1 and S2). The accuracy of the predictions is excellent in all species (98.6–99.9%). We use the common definition of accuracy as the number of correct classifications divided by the total number of classifications. The classification problem here consists of

organism	known miRNAs in data	sensitivity (miRNA loci correctly classified / miRNA loci)	specificity (non-miRNA loci correctly classified / non-miRNA loci)	prevalence (miRNA loci / (miRNA + non-miRNA loci))	accuracy (sensitivity · prevalence + specificity · (1 - prevalence))	novel miRNAs reported	novel miRNAs estimated true positives	novel miRNAs estimated true positive rate	novel miRNAs, both mature and star strands detected in a least two samples	total miRNAs detected (known and novel estimated true positives)
Human (liver)	455	79%	99.3%	2.3%	98.9%	124	75 ± 7	60 ± 6%	15	434
Human (cell lines)	631	81%	99.2%	3.2%	98.6%	147	126 ± 5	86 ± 3%	27	634
Mouse	586	82%	99.7%	1.5%	99.4%	104	85 ± 5	82 ± 5%	23	563
Sea squirt	441	71%	99.6%	0.9%	99.4%	124	55 ± 6	44 ± 5%	24	370
Fruitfly	152	78%	99.9%	0.4%	99.9%	8	1 ± 1	15 ± 15%	2	120
Nematode	154	82%	99.9%	0.4%	99.8%	32	20 ± 3	62 ± 11%	9	147
Planaria	156	88%	99.9%	0.5%	99.9%	12	11 ± 1	89 ± 4%	0	148
Anemone	40	90%	99.8%	0.2%	99.8%	28	25 ± 1	90 ± 4%	-	61

Figure 3. miRDeep2 performance on sequencing data from seven animal clades. miRDeep2 was run on Illumina sequencing data from seven animal species, representing deuterostomes (human, mouse, sea squirts), ecdysozoans (fruit fly, nematode), lophotrochozoans (planaria) and non-bilaterians (sea anemone). Accuracy is calculated as accuracy = sensitivity × prevalence + specificity (1-prevalence) and ranges from 98.6% to 99.9%. Sensitivity is calculated as the fraction of correctly classified miRNA loci. Specificity is the fraction of correctly classified non-miRNA loci. Prevalence is the fraction of analyzed loci which are miRNA loci. In the calculations, miRNA loci is set equivalent to miRBase miRNA loci, for a discussion of this assumption, see the 'Results' section. The true positive rate of novel miRNAs is estimated from the miRDeep2 built-in controls. In five of the seven species, both mature and star strand of novel miRDeep2 miRNAs were detected in at least two independent samples. No such independent detection was possible in the sea anemone data, as data from a single sample was analyzed.

distinguishing miRNA from non-miRNA genome loci. In calculating the accuracy we make the assumption that all miRNAs in miRBase represent genuine miRNA loci, while all loci that are not in miRBase are non-miRNA loci. While this assumption almost certainly does not hold in all cases, the miRBase public database arguably sets the best standard for which loci represent miRNA genes and which do not. Thus falsely annotated miRNAs in miRBase which are not reported by miRDeep2 will cause the accuracy of the algorithm to be underestimated, as will genuine novel miRDeep2 miRNAs which are not in miRBase. Conversely, the presence of genuine miRNAs which are not detected by miRDeep2 and are not in miRBase will cause an overestimation of the accuracy. We calculate accuracy from sensitivity, specificity and prevalence using the equation: accuracy = sensitivity $\times$ prevalence + specificity (1-prevalence). Sensitivity is the fraction of miRNA loci that are correctly classified by miRDeep2 and is high (71–90%) in all seven species. Specificity is the fraction of non-miRNA loci that are correctly classified and is excellent (99.2–99.9%) in all seven species. The prevalence is the fraction of loci analyzed by miRDeep2 which are miRNA. miRDeep2 reports novel miRNAs for all seven species, even though some of the species have already been heavily mined for miRNAs (human, mouse, nematode). The true positive rate as calculated by miRDeep2 built-in-controls is at least ~50% in all species except fly. Since the fly data has already been mined for miRNAs (44) we speculate that few yet remain in the data to be detected. The positive predictive value indicates the number of reported miRNAs which are genuine, summing over known and novel. We find that it ranges from 84% (sea squirt) to 99% (planaria) in the data analyzed here. In particular the positive predictive value is above 90% in all species except sea squirt, demonstrating high specificity of miRDeep2 predictions. Furthermore, in 5 of the 7 species miRDeep2 reported high-confidence miRNAs where both the mature and the star sequences were detected in at least two independent samples.

In a recent study, 108 novel mouse miRNAs were manually curated from comprehensive high-throughput sequencing data sets (38). These miRNAs are now included in the public miRBase database. Performing a computational analysis of the same data, miRDeep2 recovered 72 (66%) of these miRNA candidates. In this previous study, 17/25 (72%) novel miRNAs were validated by ectopic expression of the predicted hairpin precursor, followed by sequencing to detect the resulting Dicer processing products. The miRDeep2 sequences that overlap the tested miRNAs had a similar validation rate (79%) as the manually curated set. On top of the 108 miRNAs presented in the earlier study, miRDeep2 reported another 104 novel miRNAs from the data, of which 23 are high-confidence genes where both mature and star strands were detected in at least two independent samples.

Consistent with earlier observations (41) we note that many of the novel reported sea anemone miRNAs have read signatures that are not typical of Drosha/Dicer processing, including short loops and imprecise begin and end positions of putative mature and star strands.

miRDeep2 distinguishes miRNAs from other argonaute-bound small RNAs

The nematode data that we analyzed contains ~1.8 million 21U-RNAs, small RNAs that interact with the Prg-1 protein of the argonaute family and are involved in worm fertility. Although we did not in any of our analyses discard reads because of annotation, and although 21U-RNAs have similar length and sequence features as miRNAs (40) they only overlapped with a single miRNA candidate reported by miRDeep2 (out of 32 candidates) showing that their presence does not substantially affect the gene prediction. Similarly, the planarian data contains overall 43% piRNAs (42), small RNAs that interact with Piwi proteins of the argonaute family and normally silence transposons in the animal germline. None of the novel planarian miRNAs reported by miRDeep2 overlap with annotated piRNA clusters.

High-throughput validation of novel human miRDeep2 miRNAs

To investigate if the novel miRNAs reported by miRDeep2 are in fact dependent on the canonical biogenesis for expression, we used RNA interference to silence *Dicer* in a MCF-7 breast cancer cell line (see 'Materials and Methods' section). Small RNA libraries from unperturbed cells and from silenced cells were prepared and sequenced on the Illumina platform. The small RNAs were sampled from both cytoplasmic and total cellular fractions, however these were pooled for the unperturbed and for the silenced cells. qPCR measurements showed that *Dicer* mRNA was downregulated by 54–84% (total cellular and cytoplasmic, respectively). As negative controls we first investigated transcripts that are believed not to be frequently cleaved by Dicer. Specifically, we investigated how many sequencing reads map to snoRNA, tRNA and rRNA transcripts in the unperturbed and silenced cells (see 'Materials and Methods' section). For the snoRNA transcripts, there was a median 1% increase in the number of mapping reads following the *Dicer* silencing, corresponding to a log2 fold-change of 0.02 (Figure 4D). For the tRNA and rRNA transcripts, there was a 10% and 9% reduction, respectively (Figure 4E and 4F). For ~6300 genomic control regions that give rise to small RNAs but do not fold into hairpins, there was median reduction of 10% reads following *Dicer* silencing (Figure 4G). These results show that transcripts that are not believed to undergo frequent Dicer processing are largely unaffected by the silencing. We speculate that the small RNAs produced from these transcripts may be produced by Dicer independent pathways or degradation. As positive controls we investigated how many reads map to miRNAs from the public miRBase database. We found a median reduction of 56% corresponding to a log2 fold-change of -1.2, showing as expected that miRNA expression is substantially affected by *Dicer* silencing (Figure 4H). Taqman measurements showed that the mature miR-16 transcript was downregulated by 68–74% (total cellular and cytoplasmic fractions respectively). According to the read count miR-16 is downregulated by 40–53%, suggesting that our sequencing analysis may

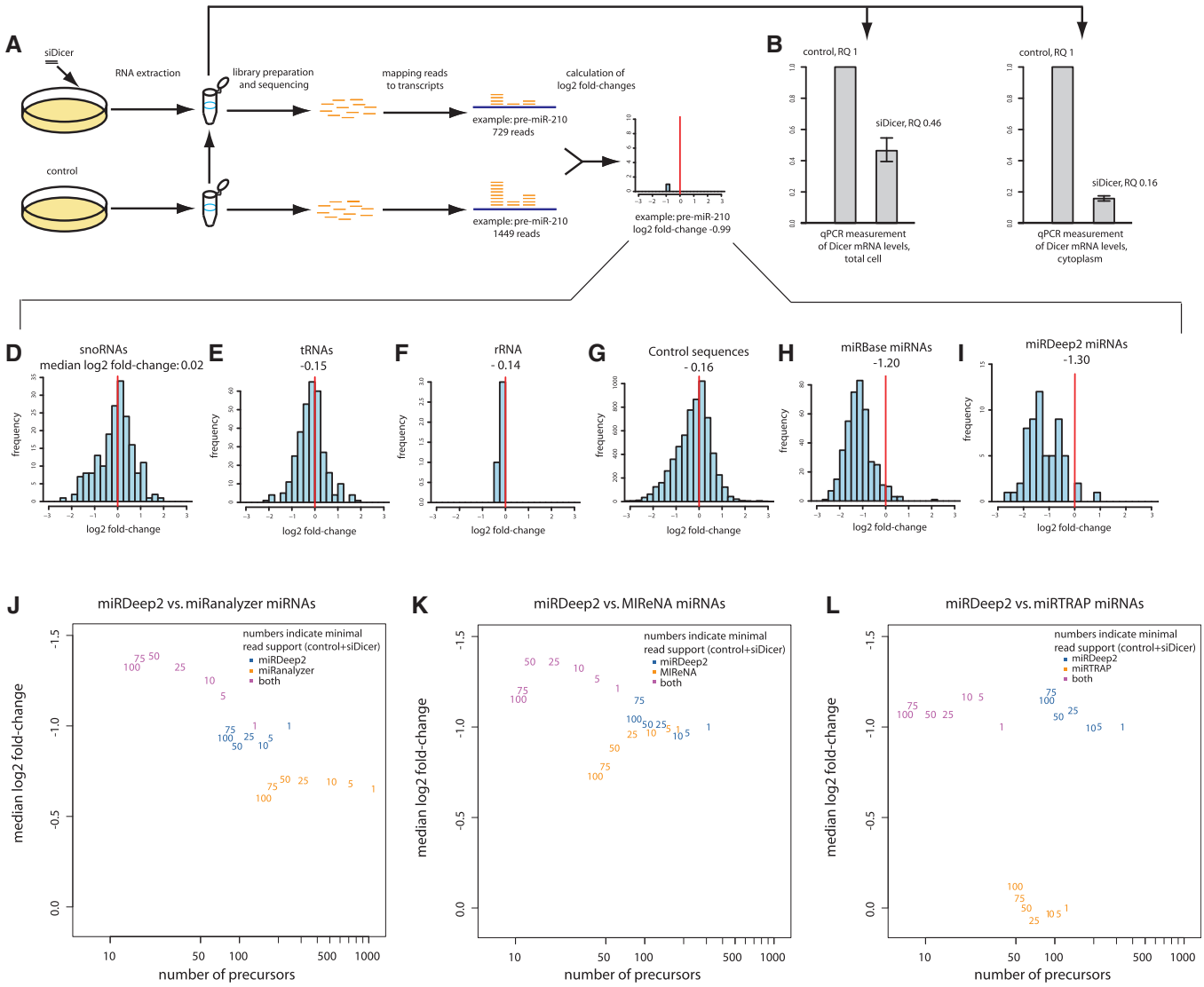

Figure 4. Effect of *Dicer* silencing on small RNA expression. RNA interference was used to silence *Dicer* in a MCF-7 cell line. (**A**) Schematic representation of the experiment; levels of *Dicer* mRNA in total cells (**B**) or cytoplasm (**C**) before and after silencing. Fold-changes in small RNA expression are noted for (**D**) snoRNAs, (**E**) tRNAs, (**F**) rRNAs, (**G**) genomic control sequences, (**H**) miRBase miRNAs, (**I**) novel miRNAs reported by miRDeep2. The median fold-change is indicated above each plot. A comparison of predictions by miRDeep2 with (**J**) miRanalyzer, (**K**) MIReNA and (**L**) miRTRAP was done. The predicted precursors are assigned to sets based on sequencing support, e.g. all the precursors in sets labeled '5' are supported by five or more sequencing reads (control + siDicer). Precursors reported only by miRDeep2 are in blue, precursors reported only by the competing program are in orange. Precursors reported by both are in purple.

underestimate the downregulation upon *Dicer* silencing. The miRNAs predicted by miRDeep2 had a log2 fold-change of −1.3 which corresponds to a median reduction of 59% upon Dicer silencing similar to those in the miRBase database (Figure 4I and Supplementary Table S1). We also predicted novel miRNAs with the competing programs miRanalyzer 0.2 (Figure 4J), MIReNA (Figure 4K) and miRTRAP (Figure 4L) taking care to use the same input data and running all programs with their default options (see 'Materials and Methods' section).

Novel miRNAs that are predicted by both miRDeep2 and the competing programs consistently show the strongest response upon the knockdown in comparison to predictions that are exclusive to miRDeep2 or the competing programs. Predictions exclusive to miRDeep2 show a stronger response than predictions exclusive to miRanalyzer or MIReNA. The miRTRAP-specific predictions only respond weakly to the *Dicer* knockdown. These results suggest that miRDeep2 performs better than competing programs for miRNA prediction. Nonetheless, the competing programs do predict miRNAs that were missed by miRDeep2 but react to the knockdown, indicating that the methods are to some degree complementary.

miRDeep2 analysis of ∼30 million RNAs consumes less than 5 h and 3 GB memory

The sequence data produced by high-throughput platforms often map to millions of genomic loci. Investigating each of

these loci can be time consuming and also unnecessary, since the vast majority have only one or two reads mapping, which in any case cannot provide solid evidence for Dicer processing. To solve this problem, miRDeep2 can be given an option to only analyze genome loci that harbor a read stack of a designated minimum height. In the default mode, miRDeep2 automatically estimates this minimum height based of the depth of the data, thus 'gearing' the analysis for the data (see 'Materials and Methods' section). In addition we have identified bottlenecks in terms of memory and time consumption and smoothed them, either by rewriting source or by incorporating better tools into the package. For instance, we have incorporated the Bowtie (34) mapping tool into all modules. Performing a full analysis on ~30 million sequenced small RNAs from human liver consumed <5 h and <3 GB memory, showing performance in the typical use case. This analysis includes mapping all reads to the human genome. In addition, we have analyzed ~100 human small RNA datasets from the SRA and GEO public databases, comprising altogether ~0.7 billion reads after stringent filtering (see 'Materials and Methods' section). miRDeep2 analyzed these data in 57 h, consuming ~4.1 GB memory. The analysis of increasingly small sampled subsets of the data suggests that analysis time increases by a factor four for every 10-fold increase in input data, while memory consumption stays nearly constant (see 'Materials and Methods' section and Supplementary Table S2). While most known miRNAs were recovered when just ~0.15 billion reads were analyzed, increasing numbers of novel miRNAs were reported with higher confidence as more data were analyzed (Supplementary Table S2).

Modularization and graphic output make miRDeep2 user friendly

In the first version of the miRDeep software package, a number of scripts had to be called consecutively by the user to run the entire analysis pipeline. In contrast, the miRDeep2 software package comprises three modules which are each run with a single command line. This means, for instance, that a raw Illumina sequencing data set can be processed, mapped to the genome (using the Mapper module), and mined for known and novel miRNAs (using the miRDeep2 module) in just two command lines. A progress report file summing up all pipeline steps is automatically generated for each analysis. The miRDeep2 module integrates all results in a .html file as well as a corresponding tab-separated file which contain detailed information on every known and novel miRNA in the data (Figure 5). In the top of the .html file is a survey of miRDeep2 performance for varying score cut-offs. For each score cut-off the sensitivity and number of true positive novel miRNAs is estimated, allowing the user to choose the sensitivity/specificity trade-off that is desirable for the particular analysis. Below is a summary table listing all known and novel miRNAs in the data. If mature and precursor miRNAs for the species analyzed have been input, the table also includes any known miRNAs that were not detected by miRDeep2. For each miRNA all generated results are given, including: miRDeep2 score, estimated probability that the reported miRNA is genuine, summary of sequences and read counts and a field that indicates if the reported miRNA matches reference rRNA and tRNA Rfam sequences. Furthermore, each miRNA is linked to graphics that provide detailed information about the secondary structure, sequencing reads aligned to the miRNA precursor as well as a miRDeep2 score breakdown (Figure 2). The summary table also links to a number of public databases, including miRBase, NCBI blast search and the UCSC genome browser for the species analyzed.

Numerous novel human miRNAs originate from diverged transposons

Our analysis of sequenced small RNAs from human liver samples yielded 124 novel miRNA candidates, while our analysis of small RNAs from human cell lines yielded 147 novel candidates (Figure 3). There is an overlap of 13 hairpins between the two sets, so in total we report 258 novel human miRNA candidates. Given that the built-in controls of miRDeep2 indicate that these are not all genuine, we conservatively curated a set of high-confidence miRNAs where both the mature and the star sequences had been detected in minimum two independent samples. This yielded a set of 42 non-redundant high-confidence miRNA candidates. A subset of 16 high-confidence miRNAs (~40%) locates to LINE, SINE or DNA transposable elements (Table 1). These elements have however all diverged from the consensus sequences (by 8–31%), suggesting that they are no longer mobile instances. While most searches for miRNA genes initially discard all candidates that locate to repeats [e.g. (15)] our results show that miRNA genes can originate from transcribed inactive transposable elements, consistent with previous findings (45,46). Several of the miRNAs locating to transposable elements were primarily supported by unambiguously mapping reads, showing that we have correctly identified the source of the small RNAs (e.g. Supplementary Figure S3). Two of the high-confidence miRNAs locate to putative protein-coding genes. One is contained in CDS and the other one overlaps with 5′ UTR. These may be false annotations or may be genuine protein-coding genes that give rise to small RNAs similar to the *Dgcr8* transcript (47,48). Two miRNAs were found to be anti-sense to the known miRNAs hsa-mir-219-2 and hsa-mir-1295, and five other miRNAs locate to annotated snoRNAs, consistent with previous reports that snoRNAs can be processed into functional small RNAs [e.g. (49,50)]. The remaining 17 predicted novel miRNAs map either to intergenic (5) or intronic (12) genomic locations.

DISCUSSION

We have tested miRDeep2 on high-throughput sequencing data from all major animal clades, including vertebrates, non-vertebrate deuterostomes, ecdysozoans, lophotrochozoans and non-bilaterians. miRDeep2 identifies miRNA genes with high accuracy (98.6–99.9%) and sensitivity (71–90%) in all clades. Furthermore, in all species

miRDeep2

Survey of miRDeep2 performance for score cut-offs 0 to 10

	novel miRNAs			known miRBase miRNAs				
miRDeep2 score	predicted by miRDeep2	estimated false positives	estimated true positives	in species	in data	detected by miRDeep2	estimated signal- to- noise	excision gearing
10	95	8 ± 3	87 ± 3 (91 ± 3%)	914	572	407 (71%)	32.3	4
9	100	8 ± 3	92 ± 3 (92 ± 3%)	914	572	408 (71%)	32.1	4
8	110	9 ± 3	101 ± 3 (92 ± 3%)	914	572	411 (72%)	32.2	4
7	119	9 ± 3	110 ± 3 (92 ± 2%)	914	572	411 (72%)	32.1	4
6	124	9 ± 3	115 ± 3 (92 ± 2%)	914	572	411 (72%)	31.3	4
5	149	11 ± 3	138 ± 3 (92 ± 2%)	914	572	454 (79%)	29	4
4	173	22 ± 4	151 ± 4 (87 ± 2%)	914	572	475 (83%)	19.9	4
3	192	58 ± 7	134 ± 7 (70 ± 4%)	914	572	478 (84%)	9.3	4
2	227	76 ± 8	151 ± 8 (67 ± 4%)	914	572	489 (85%)	7.5	4
1	335	107 ± 9	228 ± 9 (68 ± 3%)	914	572	511 (89%)	6.2	4
0	397	361 ± 17	36 ± 17 (9 ± 4%)	914	572	514 (90%)	2.2	4

Novel miRNAs predicted by miRDeep2

provisional id	miRDeep2 score	estimated probability that the miRNA candidate is a true positive	rfam alert	total read count	mature read count	loop read count	star read count	significant randfold p- value	miRBase miRNA	example miRBase miRNA with the same seed	UCSC browser	NCBI blastn	consensus mature sequence	consensus star sequence
Hs7_7976_12459	4.2e +2	0.91 ± 0.03		830	734	0	96	yes			blat	blast	ugucuuacucccucaggcacau	agugccugagggaguaagag
Hs13_24680_20135	2.8e +2	0.91 ± 0.03		566	442	0	124	yes			blat	blast	uguuguacuuuuuuuuuuguuc	acaaaaaaaaaagcccaacccu
Hs3_5769_6412	2.0e +2	0.91 ± 0.03		392	345	0	47	yes			blat	blast	caaaaacugcaauuacuuuugc	gaaaguaauugcuguuuuugcc
Hs2_5560_3658	1.8e +2	0.91 ± 0.03		365	334	0	31	yes			blat	blast	aaaaaccacaauuacuuuugc	agaaguaauugcggucuuugcc
Hs4_16510_7415	1.7e +2	0.91 ± 0.03		336	243	1	92	yes			blat	blast	caaaaacugcaguuacuuuugc	caaaagugauugcaguguuugcc
Hs13_24654_19774	1.1e +2	0.91 ± 0.03		223	151	0	72	yes		ptr- miR-548h	blat	blast	aaaaguaauugcaguuuuugc	uaaaacugcaguuauuuuugc
Hs14_26604_21006	1.0e +2	0.91 ± 0.03		205	99	0	106	yes		ptr- miR-548h	blat	blast	aaaaguaaucacuguuuuugcc	caaaaaccgugauuacuuuugc
Hs11_9141_16854	8.9e +1	0.91 ± 0.03		175	142	2	31	yes			blat	blast	uucucuauaggaagccauagca	uauguuuuccugaggagauaua
Hs10_8862_16763	7.1e +1	0.91 ± 0.03		140	135	0	5	yes			blat	blast	uugugaagaaagaaauucuuac	aagaauuucuuuuucuucacaauu
Hs13_10109_19638	6.2e +1	0.91 ± 0.03		120	63	0	57	yes			blat	blast	uaaaacccacaauuauguuugu	aaaaguaauugcgggguuugcc
HsX_11826_27786	4.9e +1	0.91 ± 0.03		88	75	0	13	yes		ppy- miR-655	blat	blast	auaauacaaccugcuaagug	cuuagcagguuguauuau
Hs1_4993_2598	4.4e +1	0.91 ± 0.03		86	64	0	22	yes			blat	blast	ucugugagaccaaagaacuacu	uuguucuuuggucuuucagcc
Hs6_7749_10863	4.1e +1	0.91 ± 0.03		80	60	0	20	yes			blat	blast	ucaggugugaaacugaggcagg	ugcucaggguugcacagcuggga
Hs11_34082_18138	3.8e +1	0.91 ± 0.03		73	37	0	36	yes			blat	blast	uaugguacuccuuaagcuaac	uuagcuuaaggaguaccagauc

Figure 5. Example miRDeep2 output: performance survey and novel human miRNAs. For each analysis, miRDeep2 outputs a single .html page that links to all results generated by the module. In the top of the .html file is a survey of miRDeep2 performance for varying score cut-offs, providing estimates of sensitivity and number of true positive novel miRNAs. Below is a table of novel miRNAs discovered in the sequencing data. Each line includes the following information on one novel miRNA candidate: the miRDeep2 score, the probability that the miRNA candidate is genuine given the evidence from the sequencing data, sequence and read count summaries, a link to a graphic representation of structure and read signature (example seen in Figure 2), a link to the UCSC genome browser for the species analyzed, and a link to NCBI blast results for the candidate precursor sequence.

except fly did miRDeep2 report numerous novel miRNAs, many of which are high-confidence candidates, where both mature and star sequence were both detected in at least two independent samples. We speculate that the fly data has already been completely mined for miRNAs.

The nematode and planarian data sets contain millions of argonaute-bound 21U-RNAs and piRNAs respectively, many of which have length and sequence features similar to miRNAs. However, even though no sequences were discarded because of annotation they did not cause significant numbers of false positives, showing that miRDeep2 can confidently distinguish miRNAs from other argonaute-bound small RNAs.

To perform high-throughput validation of our novel human miRNAs, we silenced *Dicer* and used sequencing to measure changes in small RNA expression upon the perturbation. Since normalization of small RNA data across samples is not trivial, we observed changes in negative controls (snoRNAs, rRNAs, tRNAs and non-hairpin transcripts) and positive controls (miRNAs from the miRBase public database). We observed that our novel human miRNAs have almost exactly the same response to *Dicer* silencing as do the miRNAs from the public database. miRanalyzer and MIReNA candidates were less responsive to the knockdown. In contrast, only a small subset of the miRNAs predicted by the miRTRAP algorithm responds to downregulation, indicating that many may be false positives.

We have shown that miRDeep2 can perform a full analysis of ~30 million sequenced human RNAs consuming less than 5 h and 3 GB memory. Our benchmarking shows that time consumption scales better than linear with the number of reads input, suggesting that miRDeep2 will still be efficient if output from high-throughput sequencing platforms increases by another order of magnitude. For the dataset with 740 million reads the memory consumption was ~4.1 GB which is still well within the capacities of scientific computers.

Table 1. Genomic sources of novel human miRNAs

Genomic source	Number of novel miRNAs		Details
	Cell line	Liver	
Intergenic	3	2	
Intronic	9	3	
5′-UTR	1	0	TIGD1
Coding sequence	1	0	Putative gene, clusters with hsa-miR-137
Diverged transposable elements (intronic)			
	2	0	SINE
	2	2	LINE
	6	2	DNA element: Mariner
Diverged transposable elements (intergenic)			
	0	1	LINE
	0	1	DNA element:MER91A
snoRNAs (intronic)	2	3	SnoRNA25, snoRNA36A, snoRNA33,ACA47,HBII-99B
Antisense to miRNA	1	1	hsa-miR-219-2, hsa-miR-1295
Total	27	15	

We hope that the improved interface of miRDeep2 will make the algorithm useful to a wider range of biologists who do not have extensive computational experience. The graphic output will also benefit experienced users, allowing easy manual inspection of the results and fast links to relevant databases.

In analyzing the human sequence data, we noticed that between the 124 novel miRNAs detected in liver and the 147 detected in cell lines there was only a limited overlap of 13 genes. These tendencies also hold when we limit the analysis to high-confidence miRNAs where both mature and star sequences were both detected in at least two independent samples. This suggests that human miRNA annotation is still far from saturation, consistent with the fact that novel human miRNAs are submitted to miRBase at a non-diminishing pace. However, while the number of miRNA genes steadily grows, these novel sequences account for increasingly small fractions of the cellular small RNAs. This raises the question if these very lowly abundant Drosha/Dicer products have cellular functions that are under positive selection or if they are rather under neutral selection because they are too lowly abundant to have any real effect on protein output (51). Resolving this question by looking at conservation patterns of miRNA genes and their target sites may be difficult, since lowly abundant miRNAs also tend to be less conserved. However, saturated sequencing of many closely related species combined with 'phylogenetic shadowing' (52) might reveal positive selection working on miRNAs that are only present in single animal clades. Similarly, novel high-throughput technologies like CLIP-seq (53,54) might reveal miRNA target sites that are preferentially bound by argonaute proteins, thus bypassing the need for conservation analysis. Still, in the years to come it might be much more difficult to determine if a small RNA has a function than to determine if it is a product of miRNA biogenesis.

ACCESSION NUMBER

Raw data and processed tables of the Dicer silencing experiment can be accessed by GSE31069 from the NCBI Gene Expression Omnibus at http://www.ncbi.nlm.nih.gov/geo.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Catherine Adamidi for advice on validation experiments. Xintian You helped with the analysis of data from the *Dicer* silenced cells.

FUNDING

Helmholtz-Alliance on Systems Biology (Max Delbrück Centrum Systems Biology Network to S.D.M.); the Helmholtz Association; the German Ministry of Education and Research (BMBF) and the Senate of Berlin by funds for the Berlin Institute for Medical Systems Biology (BIMSB) (315362A); China Scholarship Council (to N.L.). Funding for open access charge: Max Delbrück Centrum Systems Biology Network (MSBN).

Conflict of interest statement. None declared.

REFERENCES

1. Winter,J., Jung,S., Keller,S., Gregory,R.I. and Diederichs,S. (2009) Many roads to maturity: microRNA biogenesis pathways and their regulation. *Nat. Cell Biol.*, **11**, 228–234.
2. Chekulaeva,M. and Filipowicz,W. (2009) Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. *Curr. Opin. Cell Biol.*, **21**, 452–460.

3. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
4. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
5. Stoeckius,M., Maaskola,J., Colombo,T., Rahn,H.P., Friedlander,M.R., Li,N., Chen,W., Piano,F. and Rajewsky,N. (2009) Large-scale sorting of C. elegans embryos reveals the dynamics of small RNA expression. *Nat. Methods*, **6**, 745–751.
6. Stefani,G. and Slack,F.J. (2008) Small non-coding RNAs in animal development. *Nat. Rev. Mol. Cell Biol.*, **9**, 219–230.
7. Bushati,N. and Cohen,S.M. (2007) microRNA functions. *Annu. Rev. Cell Dev. Biol*, **23**, 175–205.
8. Lau,N.C., Lim,L.P., Weinstein,E.G. and Bartel,D.P. (2001) An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, **294**, 858–862.
9. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
10. Lee,R.C. and Ambros,V. (2001) An extensive class of small RNAs in Caenorhabditis elegans. *Science*, **294**, 862–864.
11. Friedlander,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.
12. Mathelier,A. and Carbone,A. (2010) MIReNA finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, **26**, 2226–2234.
13. Hackenberg,M., Sturm,M., Langenberger,D., Falcon-Perez,J.M. and Aransay,A.M. (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, **37**, W68–W76.
14. Hackenberg,M., Rodriguez-Ezpeleta,N. and Aransay,A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–138.
15. Hendrix,D., Levine,M. and Shi,W. (2010) miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol.*, **11**, R39.
16. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
17. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
18. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
19. Juhling,F., Morl,M., Hartmann,R.K., Sprinzl,M., Stadler,P.F. and Putz,J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.
20. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
21. Jima,D.D., Zhang,J., Jacobs,C., Richards,K.L., Dunphy,C.H., Choi,W.W., Yan Au,W., Srivastava,G., Czader,M.B., Rizzieri,D.A. *et al.* (2010) Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. *Blood*, **116**, e118–e127.
22. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
23. Somel,M., Guo,S., Fu,N., Yan,Z., Hu,H.Y., Xu,Y., Yuan,Y., Ning,Z., Hu,Y., Menzel,C. *et al.* (2010) MicroRNA, mRNA, and protein expression link development and aging in human and macaque brain. *Genome Res.*, **20**, 1207–1218.
24. Persson,H., Kvist,A., Vallon-Christersson,J., Medstrand,P., Borg,A. and Rovira,C. (2009) The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nat. Cell Biol.*, **11**, 1268–1271.
25. Mayr,C. and Bartel,D.P. (2009) Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells. *Cell*, **138**, 673–684.
26. Stark,M.S., Tyagi,S., Nancarrow,D.J., Boyle,G.M., Cook,A.L., Whiteman,D.C., Parsons,P.G., Schmidt,C., Sturm,R.A. and Hayward,N.K. (2010) Characterization of the melanoma miRNAome by Deep Sequencing. *PLoS One*, **5**, e9685.
27. Vaz,C., Ahmad,H.M., Sharma,P., Gupta,R., Kumar,L., Kulshreshtha,R. and Bhattacharya,A. (2010) Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood. *BMC Genomics*, **11**, 288.
28. Shin,C., Nam,J.W., Farh,K.K., Chiang,H.R., Shkumatava,A. and Bartel,D.P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Mol. Cell*, **38**, 789–802.
29. Taft,R.J., Simons,C., Nahkuri,S., Oey,H., Korbie,D.J., Mercer,T.R., Holst,J., Ritchie,W., Wong,J.J., Rasko,J.E. *et al.* (2010) Nuclear-localized tiny RNAs are associated with transcription initiation and splice sites in metazoans. *Nat. Struct. Mol. Biol.*, **17**, 1030–1034.
30. Liao,J.Y., Ma,L.M., Guo,Y.H., Zhang,Y.C., Zhou,H., Shao,P., Chen,Y.Q. and Qu,L.H. (2010) Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly complex subcellular distribution of miRNAs and tRNA 3′ trailers. *PLoS One*, **5**, e10563.
31. Kuchen,S., Resch,W., Yamane,A., Kuo,N., Li,Z., Chakraborty,T., Wei,L., Laurence,A., Yasuda,T., Peng,S. *et al.* (2010) Regulation of microRNA expression and abundance during lymphopoiesis. *Immunity*, **32**, 828–839.
32. Sha,A.G., Liu,J.L., Jiang,X.M., Ren,J.Z., Ma,C.H., Lei,W., Su,R.W. and Yang,Z.M. (2011) Genome-wide identification of micro-ribonucleic acids associated with human endometrial receptivity in natural and stimulated cycles by deep sequencing. *Fertil. Steril*, **96**, 150–155.
33. Hou,J., Lin,L., Zhou,W., Wang,Z., Ding,G., Dong,Q., Qin,L., Wu,X., Zheng,Y., Yang,Y. *et al.* (2011) Identification of miRNomes in human liver and hepatocellular carcinoma reveals miR-199a/b-3p as therapeutic target for hepatocellular carcinoma. *Cancer Cell*, **19**, 232–243.
34. Langmead,B. (2010) Aligning short sequencing reads with Bowtie. *Curr. Protoc. Bioinformatics*, Chapter 11, Unit 11 17.
35. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
36. Bonnet,E., Wuyts,J., Rouze,P. and Van de Peer,Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.
37. Shi,W., Hendrix,D., Levine,M. and Haley,B. (2009) A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat. Struct. Mol. Biol.*, **16**, 183–189.
38. Chiang,H.R., Schoenfeld,L.W., Ruby,J.G., Auyeung,V.C., Spies,N., Baek,D., Johnston,W.K., Russ,C., Luo,S., Babiarz,J.E. *et al.* (2010) Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev.*, **24**, 992–1009.
39. Tyler,D.M., Okamura,K., Chung,W.J., Hagen,J.W., Berezikov,E., Hannon,G.J. and Lai,E.C. (2008) Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes Dev.*, **22**, 26–36.
40. Ruby,J.G., Jan,C., Player,C., Axtell,M.J., Lee,W., Nusbaum,C., Ge,H. and Bartel,D.P. (2006) Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, **127**, 1193–1207.
41. Grimson,A., Srivastava,M., Fahey,B., Woodcroft,B.J., Chiang,H.R., King,N., Degnan,B.M., Rokhsar,D.S. and Bartel,D.P. (2008) Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, **455**, 1193–1197.
42. Friedlander,M.R., Adamidi,C., Han,T., Lebedeva,S., Isenbarger,T.A., Hirst,M., Marra,M., Nusbaum,C., Lee,W.L., Jenkin,J.C. *et al.* (2009) High-resolution profiling and discovery of planarian small RNAs. *Proc. Natl Acad. Sci. USA*, **106**, 11546–11551.
43. Kato,M., de Lencastre,A., Pincus,Z. and Slack,F.J. (2009) Dynamic expression of small non-coding RNAs, including novel

microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol.*, **10**, R54.
44. Ruby,J.G., Stark,A., Johnston,W.K., Kellis,M., Bartel,D.P. and Lai,E.C. (2007) Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res.*, **17**, 1850–1864.
45. Piriyapongsa,J. and Jordan,I.K. (2007) A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One*, **2**, e203.
46. Smalheiser,N.R. and Torvik,V.I. (2005) Mammalian microRNAs derived from genomic repeats. *Trends Genet.*, **21**, 322–326.
47. Triboulet,R., Chang,H.M., Lapierre,R.J. and Gregory,R.I. (2009) Post-transcriptional control of DGCR8 expression by the Microprocessor. *RNA*, **15**, 1005–1011.
48. Han,J., Pedersen,J.S., Kwon,S.C., Belair,C.D., Kim,Y.K., Yeom,K.H., Yang,W.Y., Haussler,D., Blelloch,R. and Kim,V.N. (2009) Posttranscriptional crossregulation between Drosha and DGCR8. *Cell*, **136**, 75–84.
49. Scott,M.S., Avolio,F., Ono,M., Lamond,A.I. and Barton,G.J. (2009) Human miRNA precursors with box H/ACA snoRNA features. *PLoS Comput. Biol.*, **5**, e1000507.
50. Ender,C., Krek,A., Friedlander,M.R., Beitzinger,M., Weinmann,L., Chen,W., Pfeffer,S., Rajewsky,N. and Meister,G. (2008) A human snoRNA with microRNA-like functions. *Mol. Cell*, **32**, 519–528.
51. Chen,K. and Rajewsky,N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat. Rev. Genet.*, **8**, 93–103.
52. Berezikov,E., Guryev,V., van de Belt,J., Wienholds,E., Plasterk,R.H. and Cuppen,E. (2005) Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, **120**, 21–24.
53. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
54. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.