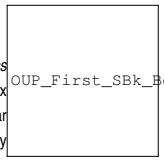
Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category



Genome analysis

miRCat2: Accurate prediction of plant and animal microRNAs from next-generation sequencing datasets

Claudia Paicu [1,2], Irina Mohorianu [2,3], Matthew Stocks [2], Ping Xu [3], Aurore Coince [3], Martina Billmeier [3], Tamas Dalmay [3], Vincent Moulton [2,*] and Simon Moxon[3,*]

[1]The Earlham Institute, Norwich Research Park, Norwich, NR4 7UG, United Kingdom
[2]School of Computing Sciences and
[3]School of Biological Sciences, University of East Anglia, Norwich Research Park, NR4 7TJ, Norwich, United Kingdom

*To whom correspondence should be addressed

Associate Editor: Dr. Inanc Birol

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: MicroRNAs are a class of ∼21-22 nucleotide small RNAs which are excised from a stable hairpin-like secondary structure. They have important gene regulatory functions and are involved in many pathways including developmental timing, organogenesis and development in eukaryotes. There are several computational tools for miRNA detection from next-generation sequencing (NGS) datasets. However, many of these tools suffer from high false positive and false negative rates. Here we present a novel miRNA prediction algorithm, miRCat2. miRCat2 incorporates a new entropy-based approach to detect miRNA loci, which is designed to cope with the high sequencing depth of current NGS datasets. It has a user-friendly interface and produces graphical representations of the hairpin structure and plots depicting the alignment of sequences on the secondary structure.
Results: We tested miRCat2 on a number of animal and plant datasets and present a comparative analysis with miRCat, miRDeep2, miRPlant and miReap. We also use mutants in the miRNA biogenesis pathway to evaluate the predictions of these tools. Results indicate that miRCat2 has an improved accuracy compared with other methods tested. Moreover, miRCat2 predicts several new miRNAs that are differentially expressed in wildtype versus mutants in the miRNA biogenesis pathway.
Availability: miRCat2 is part of the UEA small RNA Workbench and is freely available from http://srna-workbench.cmp.uea.ac.uk/.
Contact: v.moulton@uea.ac.uk, s.moxon@uea.ac.uk.
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

MicroRNAs (miRNAs) are a class of small non-coding RNAs (sRNAs) that are excised from a hairpin-like secondary structure of a primary transcript (Bartel, 2004; Kim, 2005). They are present and functional in metazoa and in some viruses; their mode of action consists of the down regulation of the target gene(s) through post transcriptional silencing (Bartel, 2004; Kim, 2005; Chen, 2005). The identification and characterization of miRNAs, which are ∼21-22nt in length, has developed as a major research topic due to their important role in gene regulation and influence on pathways such as hematopoiesis, apoptosis, cell proliferation and tumorgenesis (Cheng *et al.*, 2005; Iorio *et al.*, 2005; Esquela-Kerscher and Slack, 2006; Jones-Rhoades *et al.*, 2006; Lu *et al.*, 2008; Pérez-Quintero *et al.*, 2010).

1.1 miRNA biogenesis and function

In animals, miRNA genes are transcribed by RNA polymerase II to generate long capped and polyadenylated transcripts (termed pri-miRNAs) (Lee *et al.*, 2004; Cai *et al.*, 2004; Kim, 2005; Xie *et al.*, 2015). The Drosha protein recognises the hairpin structure of the pri-miRNA and initiates the first processing step ('cropping') (Lee *et al.*, 2003; Han *et al.*, 2004; Denli *et al.*, 2004; Gregory *et al.*, 2004; Kim, 2005; Zeng *et al.*, 2005). The product of this nuclear processing step is a ∼70-nucleotide (nt) precursor (pre-miRNA), which folds into a short stem-loop structure with a ∼2-nucleotide 3' overhang (Kim, 2005). A nuclear export factor (Exportin-5) recognises this structure as a signature motif and exports it from the nucleus to the cytoplasm (Yi *et al.*, 2003; Lund *et al.*, 2004; Bohnsack *et al.*, 2004; Bartel, 2004; Kim, 2005). Here, a Dicer protein removes the loop-region and gives rise to the miRNA duplex (process known as 'dicing') (Bernstein *et al.*, 2001; Grishok *et al.*, 2001; Hutvágner *et al.*, 2001; Ketting *et al.*, 2001; Bartel, 2004; Kim, 2005; Ha and Kim, 2014). The duplex is then separated and usually one strand is selected as the mature miRNA, whereas the other strand may be degraded; in some cases both 3' and 5' miRNAs are stable and functional (Khvorova *et al.*, 2003; Ha and Kim, 2014).

The biogenesis of miRNAs in plants is similar to that of animals, with some differences. While in animals the length and structure of the pre-miRNA hairpin is fairly consistent, in plants it is longer and much more variable (100-300nt) (Cuperus *et al.*, 2011). A DICER-LIKE1 (DCL1) enzyme excises the miRNA duplex from the pri-miRNA, in the nucleus (Reinhart *et al.*, 2002; Park *et al.*, 2002; Papp *et al.*, 2003; Kurihara and Watanabe, 2004; Xie *et al.*, 2004), then the small RNA methyltransferase hua enhancer1 (HEN1) adds a methyl group to the 3' end to stabilise it (Yu *et al.*, 2005; Xie *et al.*, 2015). The duplex is then transported from the nucleus to the cytoplasm by hasty (HST), a homolog of exportin 5 (Chen, 2005; Xie *et al.*, 2015). The duplex is then separated in the cytoplasm, giving rise to the mature miRNAs (Chen, 2005; Xie *et al.*, 2015).

The mature miRNA is incorporated into the RNA-induced silencing complex (RISC) (Bartel, 2004; Eamens *et al.*, 2009; Wu *et al.*, 2009; Fabian and Sonenberg, 2012), where it is bound by AGO proteins and guides the complex to complementary messenger RNA (mRNA) sequences (usually within the 3' UTR, in animals, and within the coding region, in plants) (Bartel, 2009; Djuranovic *et al.*, 2012; Bazzini *et al.*, 2012; Ameres and Zamore, 2013). miRNAs can regulate critical cellular and developmental processes (Cheng *et al.*, 2005; Iorio *et al.*, 2005; Esquela-Kerscher and Slack, 2006; Lu *et al.*, 2008; Ameres and Zamore, 2013). In plants, miRNAs are also involved in diverse responses to stresses such as drought, salt, cold, oxidative, nutrient deficiency as well as biotic stresses (Jones-Rhoades *et al.*, 2006; Pérez-Quintero *et al.*, 2010; Xie *et al.*, 2015).

1.2 Computational detection of miRNAs

Over the last decade, various computational tools have been developed for identifying miRNAs from next-generation sequencing (NGS) datasets, using features of the miRNA biogenesis. Some of the more commonly used tools, in temporal order of appearance, are: miRDeep (Friedländer *et al.*, 2008), miRCat (Moxon *et al.*, 2008), miReap (http://mireap.sourceforge.net/), MIReNA (Mathelier and Carbone, 2010), miRAnalyzer (Hackenberg *et al.*, 2009), miRDeep-P (Yang and Li, 2011), miRDeep2 (Friedländer *et al.*, 2012), MaturePred (Xuan *et al.*, 2011), miRDeep* (An *et al.*, 2013), miRAuto (Lee *et al.*, 2013), miRPlant (An *et al.*, 2014), miR-PREFeR (Lei and Sun, 2014), Mirinho (Higashi *et al.*, 2015) and miRA (Evers *et al.*, 2015). Many of these approaches, including the miRCat tool, suffer from high false positive and false negative rates and also lack of consistency across species (Li *et al.*, 2012; Williamson *et al.*, 2012; Kang and Friedländer, 2015).

The miRCat algorithm groups reads on proximity on the reference genome. It then selects one candidate from each locus and computes discriminative features on their secondary structure, to classify them as miRNAs. miRCat was introduced when NGS sequencing depth was typically orders of magnitude smaller compared to current NGS datasets. The higher sequencing depth strongly influences the grouping approach which can result in high false positive and high negative rates (Mohorianu *et al.*, 2013). Sequencing depth is also problematic for many of the tools mentioned above, for similar reasons (Tucker *et al.*, 2009; Baker, 2010).

To overcome this, we have developed miRCat2, a new miRNA prediction tool, applicable on both animal and plant datasets, which incorporates elements of its predecessor miRCat (Moxon *et al.*, 2008), and discriminative features from miRDeep2 (Friedländer *et al.*, 2012). miRCat2 implements a new approach to differentiate miRNA candidates from background sequences, then applies novel filters on the candidate sequence alignments and secondary structure. miRCat2 is part of the UEA small RNA Workbench (Stocks *et al.*, 2012) and it has a user-friendly interface, as well as a command-line option, which allows the integration into bioinformatics workflows. The algorithm is performing well on animal datasets; it also allows the detection of complex structures and even multiple miRNA loci within a single precursor in plants.

To assess the performance of miRCat2, we have compared it to miRCat (Moxon *et al.*, 2008), miRDeep2 (animal data) (Friedländer *et al.*, 2008), miRPlant (plant data) (An *et al.*, 2014) and miReap (http://mireap.sourceforge.net/). We have chosen these tools based on their popularity and on benchmarking results (Li *et al.*, 2012; Williamson *et al.*, 2012; Kang and Friedländer, 2015), which, to our best knowledge, classify them as generally the most advantageous methods. The comparison is performed on a number of animal and plant datasets; we also used mutants which affect the miRNA biogenesis pathway to validate the predictions.

2 METHODS

2.1 Overview

In Supplementary file SD1, Figure SD1.1 we present an overview of the workflow for the miRCat2 algorithm. After mapping the reads to the reference genome (using PatMaN (Prüfer *et al.*, 2008), full length, with 0 gaps and 0 mismatches), the algorithm selects reads based on abundance, then filters on read alignment patterns and secondary structure of the putative pre-miRNA hairpin. We now present a detailed description of the algorithm.

2.1.1 Selecting candidates

miRCat2 implements a method of candidate selection designed to deal with high depth datasets. As sequencing depth increases, degradation products may obscure miRNA peaks (see Supplementary file SD1, Figure SD1.2). To cope with this, we focus on selecting all the peaks at any given genomic location, while discounting sequences with abundances at or below a background level that we compute from the data. It is known that mature miRNAs and their complementary miRNA* sequence generally have higher abundances than non-miRNAs (Lau *et al.*, 2001). When aligning miRNA reads back to the pre-miRNA locus we see characteristic peaks forming, corresponding to the 5' and 3' miRNA sequences (Supplementary file SD1, Figure SD1.2,A). We can use this information to select a restricted group of sequences as candidates, on which further analysis is performed.

To identify putative miRNA loci based on "peaks" of abundance, we use the following procedure:

a) The genome is split into consecutive windows of size l_w nucleotides (nts), with an overlap of l_o nts (Mohorianu *et al.*, 2013);

b) Each window is split into subwindows of size l_{sw} and the mapped reads are assigned to subwindows based on location;

c) Each window is compared with a random uniform distribution (RUD) on subwindow abundances, using the entropy-based Kullback-Leibler

divergence (KLD) (Kullback and Leibler, 1951; Mohorianu *et al.*, 2011), using: $D_{\mathrm{KL}}(P\|Q) = \sum_i |\ln\left(\frac{P(i)}{Q(i)}\right)|P(i)$, where i is the index of the subwindow, Q is the RUD and P is the abundance distribution on the current window. The probabilities for each subwindow are calculated from the read abundances: $P_{(i)} = \frac{\sum r_{sw}}{\sum r_w}$, where r_{sw} represents the abundance of the reads mapping to the subwindow and r_w represents the abundance of the reads mapping to the window, after a default offset of 1 has been added to each subwindow, to avoid reads with low expression to be classified as peaks. The probability for the RUD is calculated using the following equation: $Q_{(i)} = \frac{1}{\#_{sw}}$, where $\#_{sw}$ represents the total number of subwindows contained in a window;

d) A KLD score close to 0 indicates a uniform distribution i.e. no peak is present. If the distribution is a RUD, then it is unlikely that a miRNA is present at the given location and the window is discarded. If the KLD is greater than a threshold (rud_val, empirically determined), then the current window contains at least one peak (the method can also detect multiple peaks). In this case, the subwindow with the highest peak is identified and the most abundant sRNA is selected. The KLD is applied again on a restricted area around this sRNA ($plateau\_range$) to avoid detecting a peak that is actually a plateau (multiple neighbouring subwindows that are all highly expressed). If this filter is passed, the sRNA is removed from the distribution and saved as a miRNA candidate for further investigation;

e) The KLD is recalculated with the new distribution. If the new KLD is still greater than the threshold, steps c) to e) are repeated until we reach an RUD (no more peaks). All removed sRNAs are miRNA candidates and are analysed using the following steps.

2.1.2 Filtering the sequences

After miRNA candidates are selected, potential false-positive predictions are excluded from down-stream analysis using a rule-based approach. First, we discard the sequences that map to the genome more than $repeats$ times as high-confidence miRNAs are unlikely to be derived from repetitive regions of the genome (Meyers *et al.*, 2008; Kozomara and Griffiths-Jones, 2014) (user-configurable parameter).

Second, a size class distribution filter is applied, allowing us to focus on reads between 21 to 23 nt, which is the expected miRNA range. To check if the miRNA candidates are within the range, we compute the KLD on size classes, comparing the sRNA size class distribution (P) to a RUD on all size classes (Q) (Mohorianu *et al.*, 2013). The sequences contributing to the sRNA size class distribution are all the reads incident to the putative miRNA precursor. If the KLD result is $> rud\_val$, then the size class distribution is different from random. We investigate whether the most abundant size class falls between 21 to 23 nts, otherwise the sRNA locus is discarded. Since a small set of annotated miRNAs in miRBase fall outside of this size range these values are configurable (min_len, max_len).

Third, to check whether the candidates have a miRNA-like alignment of incident reads, we also apply a filter that selects sequences with evidence of precise processing of the pre-miRNA by Drosha (animals) and Dicer (plants and animals) (Bartel, 2004; Kim, 2005; Chen, 2005) i.e. the presence of one or two peaks corresponding to the miRNA/miRNA*. This filtering step ensures that the majority of reads aligned to the miRNA/miRNA* location have a high overlap (are variants of each other), and have the same genomic orientation. The distribution of reads of a genuine miRNA should have a similar shape to that shown in Supplementary file SD1, Figure SD1.2,A compared to a locus generated from random RNA degradation, Figure SD1.2,B.

We define a cluster as all sequences that map to the same genomic location, having the start and the end of the mapping position within $clear\_cut$ nts of each other. The algorithm for the classification of clusters is presented in Supplementary file SD2. We identify all clusters on the window corresponding to each selected miRNA candidate, s; next, to evaluate the existence of a precise excision (e.g. resulting from Drosha and/or Dicer cleavage), we use the following criteria: (i) if the sum of the abundances of all sequences with same start and end positions ($\pm clear\_cut$ nts) as s represent $clear\_cut\_percent\%$ of the total abundance of the cluster, then s is kept for subsequent analysis; otherwise, it is discarded; (ii) if the sum of the abundances of all sRNAs from adjacent clusters that overlap with s with more than $clear\_cut$ nts represents less than $overlap\_percent\%$ of the total abundance of the s cluster, then s is kept for further analysis; otherwise, it is discarded.

2.1.3 Using the secondary structure to determine the candidate pre-miRNA

Most methods for miRNA prediction extract a fixed, arbitrary flanking region containing the miRNA candidate and fold it using RNA secondary structure prediction tools (Lorenz *et al.*, 2011) to identify a suitable hairpin-like precursor (Moxon *et al.*, 2008; Friedländer *et al.*, 2012; An *et al.*, 2014). However, this approach is highly dependent on the length of the flanking region; therefore choosing an optimal length is a critical step. To address this, we employ RNALfold (Lorenz *et al.*, 2011), previously used by miR-PREFeR (Lei and Sun, 2014) and miRA (Evers *et al.*, 2015), which folds a large window giving all possible structures contained within that region. To detect the most appropriate secondary structure, we consider a window of max_fold_len nts on each side of the miRNA candidate, ensuring that it is wide enough to capture the pre-miRNA structure.

RNALfold outputs a list containing all possible secondary structures for the selected region, in dot-bracket notation, and their corresponding minimum free energies (MFE). To compare the stability of two subwindows of differing lengths, we calculate the adjusted minimum free energy (aMFE), per 100 nt, for each secondary structure, as follows: $aMFE = \frac{MFE}{fold\_length} * 100$. The secondary structures that contain the miRNA candidate are kept for subsequent filtering which includes the evaluation of the hairpin length; maximum aMFE; and features specific to the hairpin structure (full details of parameters are listed in Supplementary file SD3). If there is more than one subwindow whose secondary structure passes all filters, the one with the lowest aMFE is accepted as the true precursor.

miRCat2 computes a score for the proposed precursor calculated based on the miRDeep2 model, as described in (Friedländer *et al.*, 2008). The score indicates the strength of the prediction, but it does not influence the output of the method. It could be used as a ranking criteria for the results, a higher score meaning the prediction has a higher probability of being a true miRNA.

2.2 Implementation

The miRCat2 algorithm is part of the UEA small RNA Workbench (Stocks *et al.*, 2012) and is written in Java, version 1.8+; for optimal results, we recommend using the latest, stable, Java version. It can run on any operating system (Windows, Linux, Mac OSX). In addition it can be executed either through the user-friendly interface or from the command line. Two sets of default parameters are provided, one for animals and one for plants, although the user can adjust these parameters. The default parameters were set according to rules generally applicable to the annotated miRNAs from miRBase (Kozomara and Griffiths-Jones, 2014) for each specific Kingdom. A list of all parameters and their default values is presented in Supplementary file SD3.

miRCat2 requires as input a reference genome and a set of sRNA sequencing data (fasta format, non-redundant, with the adaptors trimmed). The files can be processed from fastq to the necessary format using the UEA small RNA Workbench (Stocks *et al.*, 2012). The environment can

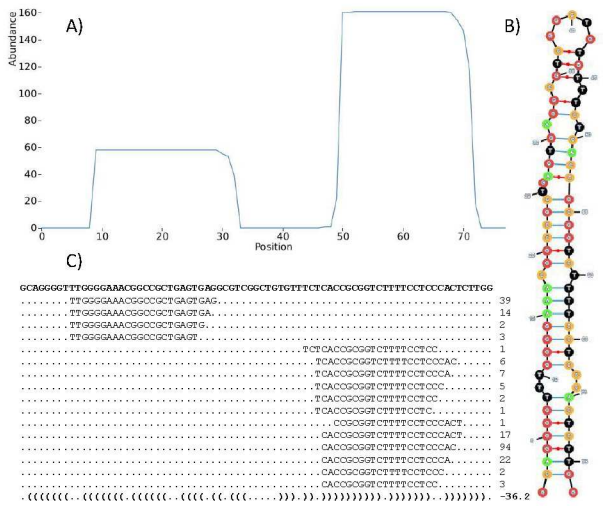


Fig. 1. Output of miRCat2 for a predicted sequence corresponding to hsa-mir-2110 (chromosome 10), depicting A) precursor presence plots, B) precursor secondary structure and C) alignment of incident reads. A) On the x-axis we represent each position along the miRNA hairpin; on the y-axis we represent the point abundance calculated as the algebraic sum of the abundances of incident reads. B) Precursor secondary structure, colour-coded for each nucleotide type (A - green, C - orange, G - red, T - black). C) Alignment of incident reads on the precursor; the numbers of the right represent the raw read abundance. The last line presents the secondary structure in dot-bracket notation, together with its MFE.

also be used to map the reads to the reference genome using PatMaN (Prüfer *et al.*, 2008), full length, with 0 gaps and no mismatches. The sequences not mapping to the reference genome are discarded.

The output of miRCat2 is presented as (i) a PatMaN file, containing the predicted miRNA coordinates, (ii) a csv file, containing additional information about the miRNA*, hairpin and existing miRNA annotations, (iii) a PDF file including, for every predicted miRNA precursor, coverage plots of mapped abundances; and (iv) a text file containing, for every prediction, the read alignments on the precursor (Figure 1).

miRCat2 uses RNALfold from the ViennaRNA package (Lorenz *et al.*, 2011) for detecting the secondary structure, randfold (Bonnet *et al.*, 2004) for calculating the statistical significance of the precursor structure. All dependencies are included in the download package and no extra installation is required. The code can be downloaded from http://srna-workbench.cmp.uea.ac.uk/downloadspage/, where users can also find the documentation and example files.

2.3 Fold change computation

To validate miRNA predictions, we estimate fold changes between wild type and mutants in the miRNA biogenesis pathway. To do this, we consider only the genome mapping reads. To compare datasets with different sequencing depths, we normalize all abundances using the reads per million (RPM) method (Mortazavi *et al.*, 2008) to the median total count (MTC) of each experiment (McCormick *et al.*, 2011; Dillies *et al.*, 2013). The method used for calculating the fold changes between wild type and mutants in the miRNA biogenesis pathway is presented in Supplementary file SD2.

3 MATERIALS

3.1 Data

To assess the performance of miRCat2, we ran it on multiple organisms and benchmarked the results against other commonly used miRNA detection tools, miRCat (version srna-workbenchV3.2), miRDeep2 (version miRDeep2.0.0.7), miRPlant (version miRPlant_V5) and miReap (version mireap_0.2). The organisms we considered are *D. rerio* (Cifuentes *et al.*, 2010), *H. sapiens* (Vaz *et al.*, 2010; Shin *et al.*, 2010; Somel *et al.*, 2010; Hou *et al.*, 2011; Friedländer *et al.*, 2014; Kim *et al.*, 2016), *M. musculus* (Bosson *et al.*, 2014; Noh *et al.*, 2014; Groenendyk *et al.*, 2014*b*,*a*; Modzelewski *et al.*, 2015; Meng *et al.*, 2015), *C. elegans* (Garcia-Segura *et al.*, 2015), *D. melanogaster* (Lee *et al.*, 2014), *H. melpomene* (Surridge *et al.*, 2011), *X. laevis* (Ahmed *et al.*, 2015) (animal datasets), *S. lycopersicum* (Lopez-Gomollon *et al.*, 2012; Kravchik *et al.*, 2014), *G. max* (Curtin *et al.*, 2016) and *A. thaliana* (Wang *et al.*, 2011) (plant datasets). We have downloaded these datasets from the GEO (Barrett *et al.*, 2013) and SRA (Leinonen *et al.*, 2010) databases. We also generated an *A. thaliana* dataset, as described in Supplementary file SD2. A description of the data processing can be found in Supplementary file SD2. Information about the genomes used, accession numbers of small RNA datasets, trimmed adapter sequences and number of reads in each dataset can be found in Supplementary file SD4.

4 RESULTS

To evaluate the performance of miRCat2, miRCat, miRDeep2, miRPlant and miReap, we generated the miRNA predictions, using all tools, on the same input datasets. We filtered the output of each tool as recommended by their authors (miRCat2: no filtering, miRCat: no filtering, miRDeep: filter by score cut-off of 0, miRPlant: filter by score cut-off of 4, miReap: no filtering). For each method and input set we determined the average number

Table 1. Performance comparison of benchmarked tools. miRCat2 performs well consistently, with a good specificity and sensitivity trade-off, while miRCat and miReap struggle in terms of specificity, especially in plants. miRDeep2/miRPlant have good specificity, but lack in sensitivity.

Animals

Organism	Tool	High-conf. miRNAs	Low-conf. miRNAs	Novel predictions	Specificity (%)	Sensitivity (%)
H. sapiens (23 datasets)	miRCat2	159	83	72	78.6 (±9.1)	30.6 (±3.3)
	miRCat	122	67	27	87.9 (±5.8)	23.9 (±2.5)
	miRDeep2	149	61	14	94 (±2.7)	26.5 (±4.5)
	miReap	148	108	227	52.3 (±14.3)	32.5 (±7.4)
M. musculus (21 datasets)	miRCat2	147	25	23	90.5 (±7.5)	39.8 (±3.2)
	miRCat	124	20	20	88.5 (±8.3)	33.5 (±1.9)
	miRDeep2	117	14	2	98.6 (±2)	29.7 (±7.2)
	miReap	114	21	134	48.7 (±12.3)	31.6 (±8.5)
D. rerio (2 datasets)	miRCat2	141	145	42	93.6 (±2.4)	88.6 (±2.3)
	miRCat	101	88	26	87.9 (±0.3)	58.2 (±2.5)
	miRDeep2	120	111	27	89.7 (±1.3)	71.5 (±3.0)
	miReap	137	132	43	86.2 (±0.2)	82.9 (±0.2)

Plants

Organism	Tool	High-conf. miRNAs	Low-conf. miRNAs	Novel predictions	Specificity (%)	Sensitivity (%)
A. thaliana (7 datasets)	miRCat2	66	44	8	93.6 (±2.7)	38.3 (±2.7)
	miRCat	51	57	167	40.9 (±9)	37.9 (±1.8)
	miRPlant	62	52	7	93.3 (±5.4)	39.3 (±14.9)
	miReap	6	8	121	14.5 (±8.5)	4.9 (±0.6)
S. lycopersicum (14 datasets)	miRCat2	15	13	233	11.6 (±5)	44.2 (±12.8)
	miRCat	14	16	1204	2.7 (±1.1)	48 (±4.8)
	miRPlant	11	7	45	30.3 (±7)	28.9 (±13.1)
	miReap	4	5	1619	0.7 (±0.3)	13.6 (±3.2)
G. max (2 datasets)	miRCat2	N/A	129	269	32.7 (±3.8)	34.9 (±1.1)
	miRCat	N/A	149	865	15.4 (±4.5)	40.2 (±0.8)
	miRPlant	N/A	80	74	52 (±0.7)	21.6 (±4.9)
	miReap	N/A	25	2243	1.2 (±0.3)	6.8 (±0.8)

of high-confidence and low-confidence miRNA precursors from miRBase v21 (Kozomara and Griffiths-Jones, 2014), the average number of new miRNA predictions, average specificity (percentage of miRBase annotated miRNAs within the output) and average sensitivity rates (percentage of miRNAs detected out of the total number of miRNAs expressed in the sample file). The averages for each organism are presented in Table 1 (full results for each dataset are described in Supplementary file SD5). We used miRBase as a reference of accepted/studied miRNAs, although we acknowledge its caveats (Saçar *et al.*, 2013).

To calculate the sensitivity, any miRNA precursor with at least one incident read, was considered to be expressed in the given sample. This approach includes low abundance miRNAs, which may be difficult to predict, resulting in overall low sensitivity.

Comparing the prediction accuracy of miRCat2 with miRCat and miRDeep2/miRPlant, we observe that miRCat2 has comparable specificity to other methods, whilst achieving an improved sensitivity. In particular, we detect a higher number of known miRNAs, whilst avoiding the proportional increase in the number of new miRNAs predictions. For example, in *M. musculus*, miRCat2 detects 41 more miRNAs than miRDeep2, which has the highest specificity, while predicting only 21 additional (potentially new) miRNAs. Moreover, miRCat2 predicts the highest number of high-confidence miRBase miRNAs in all tests. For the *H. sapiens* samples, we see that miReap predicts more known miRNAs, although at a cost to specificity, since it generates a large number of new predictions (155 more than miRCat2), which may be false positives. In all other organisms miReap performs poorly, especially in plants, where both sensitivity and specificity are low.

To validate the miRCat2 predictions, we investigated whether the predicted mature miRNAs were dependent on Dicer/DCL1, Drosha and DGCR8 processing, known to be key factors in miRNA biogenesis in plants and animals, respectively. We expect *bona-fide* miRNAs to have reduced expression in Dicer, Drosha, DGCR8 knock-out or knock-down versus wildtype samples. We consider a predicted miRNA as being down-regulated in the mutant samples if the normalized expression is at least two fold lower in the mutant, when compared with the wild type.

To evaluate the quality of the datasets, we produced sample vs sample scatter plots using the normalized expression levels in wild type and mutant samples for miRBase miRNAs (see Supplementary file SD1, Figure SD1.4). If the mutation was successful, we expect to see higher counts in the wildtype than in the mutant samples, therefore the plots should show a shift of the points above the diagonal; this pattern can be observed in the majority of cases. However, for *D. rerio* the pattern is not very clear; also in *G. max* and *M. musculus* the points are grouped on the diagonal. Nevertheless, in all cases more than a half of the points are situated above the diagonal. This suggests that these datasets contain overall lower percentages of differentially expressed miRNAs, and this is reflected in the cumulative plots too. Note that in the *H. sapiens* wildtype vs. Drosha mutant, there are some miRNAs that are located below the diagonal (more highly expressed in the mutant). This is probably because they have a Drosha-independent biogenesis pathway and therefore appear to be more highly expressed in the mutant (Kim *et al.*, 2016).

In Supplementary file SD1, Figure SD1.3 we compare the performance of miRCat2, miRCat, miRDeep, miReap and miRPlant with and without filtering. For miRCat2, we used a score cut-off of five (empirically observed to separate most new predictions from conserved miRNAs). The filtering has some impact both on miRCat2 and miRDeep2 in *H. sapiens*. In plants however we observe that miRCat2 performs well irrespective of this filtering, with a particularly large impact for miRPlant. For comparability purposes, we computed the cumulative plots of log_2 fold changes only on unfiltered outputs (see Figure 2).

For tools with high prediction accuracy we expect to see a significant differential expression (downregulation in the mutant samples) for the majority of the predicted miRNAs. As a control dataset containing reads independent in the miRNA biogenesis pathway, we use RFAM tRNA and snoRNA transcripts. As expected, their expression level is not decreased in the mutant samples; moreover, in the animal datasets the expression of these transcripts is upregulated, due to the stochasticity of the sequencing technology. In plant samples we observe little differential expression for the control sequences, as the biogenesis of plant sRNAs is more complex. All tools produce a substantially different cumulative differential expression curve compared to the control dataset; miRCat2 performs better than other tools in all but one of the experiments.

In the *H. sapiens* vs. Dicer knock-out sample (see Figure 2A), we observe that miRCat2 is a close second to miRCat, while in plant datasets there is a substantial gap between miRCat2 and the other tools, supporting the improved accuracy of miRCat2. For *S. lycopersicum*, miRCat2 shows a low specificity when detecting annotated miRNAs due to a low number of entries for this species (77 annotated precursors). However, the cumulative plots indicate that the new predictions are likely true miRNAs which have not yet been annotated in *S. lycopersicum* (54 out of the 190 new predictions are orthologues of plant miRNAs; see Supplementary file SD6 for details).

Next, we produced cumulative plots on the differential expression frequency only for the sequences that were not previously included in miRBase and therefore are potential new miRNAs (see Supplementary file SD1, Figure SD1.5). This subset contains a high proportion of putative miRNAs downregulated in the mutant samples, although to a lesser extent than the sequences included in the plots for all predictions. We observe no change in the ranking of the tools, miRCat2 performing better than the other tools in each of the experiments. In *M. musculus* we observe a decrease in the percentage of sequences with at least a 2-fold change in all tools, due to the low number of new predictions. The high percentage of differentially expressed sequences among new predictions, especially in plants, indicate that these sequences are likely to be *bona-fide* miRNAs.

To evaluate the low overall sensitivity rates, we created cumulative plots using as input the miRNAs present in the datasets, but not detected by each tool. We expect these annotated miRNAs to have low counts in

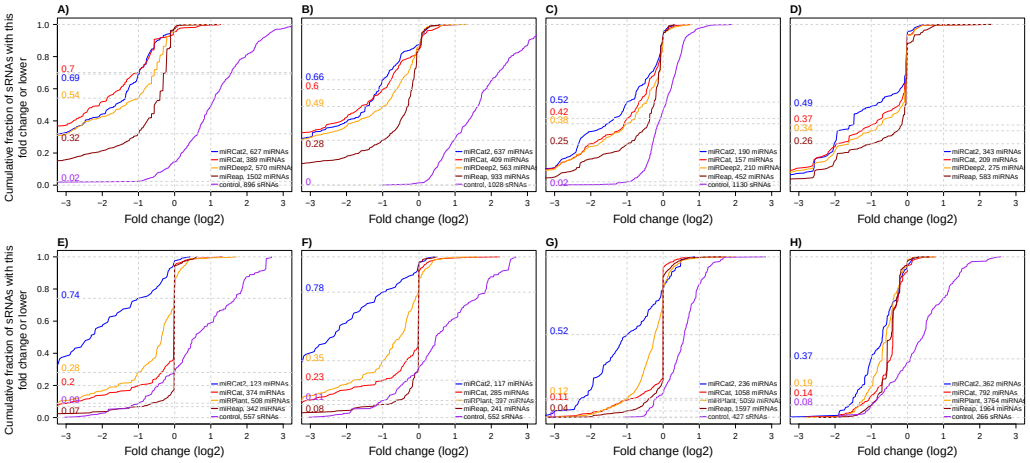


Fig. 2. Cumulative plots of log_2 fold changes of control vs. mutant datasets, calculated on the output of miRCat2, miRCat, miRDeep2/miRPlant and miReap and a control dataset formed of tRNAs and snoRNAs. We present results for H. sapiens (subplots A) Dicer and B) Drosha knock-out), M. musculus (subplot C)), D. rerio (subplot D)), A. thaliana (subplots E) and F)), S. lycopersicum (subplot G)) and G. max (subplot H)). miRCat2 has the highest percentage of DE miRNAs in all but one of the experiments, were it classifies as a close second to miRCat. A) H. sapiens wildtype vs. Dicer knock-out. B) H. sapiens wildtype vs. DROSHA knock-out. C) M. musculus wildtype vs. DGCR8 knock-out. D) D. rerio wildtype vs. Dicer knock-out. (E-F) A. thaliana wildtype vs. Dicer knock-down. G) S. lycopersicum wildtype vs. DCL1 knock-down. H) G. max wildtype vs. DCL1 knock-down.

the input samples. Some of these sequences could also be misannotations in miRBase and exhibit features not consistent with canonical miRNA structure and biogenesis. As a result, their expression would not be affected in the mutant samples. Consequently, we expect to see a smaller differential expression between the wildtype and mutant samples in the cumulative plot i.e. a curve closer to the control line. In Supplementary file SD1, Figure SD1.6, we observe a clear change in the shape of the cumulative plots for each tool (especially for miRCat2), suggesting that these miRNAs might not present the canonical miRNA features or were lowly expressed in the datasets analysed. Also, it is notable that miRCat2 consistently performs well, suggesting that it is less prone to false positives than other methods.

All miRCat2 new predictions are given in Supplementary file SD7.

5 DISCUSSION

We presented a new tool for miRNA prediction, miRCat2, applicable on both plant and animal data, which can be run both from the UEA small RNA Workbench graphical interface and from the command line.

We tested miRCat2 on ten model organisms and compared its results with four commonly used tools for miRNA discovery (miRCat, miRDeep2, miRPlant and miReap). miRCat2 shows a good trade-off between sensitivity and specificity (relative to miRBase annotation), performing well in both metrics, while other tools generally performed well only for one of these measures. More specifically, miRDeep2 and miRPlant had good specificity rates, but lacked in sensitivity (annotated miRNAs are not predicted). miReap had a good sensitivity in animals, but lacked in specificity, allowing a high number of new predictions, which could potentially contain false positives.

To evaluate the accuracy of the predictions we used the miRBase annotations and the objective and biologically meaningful mutant test (using Dicer/DCL1, Drosha and DGCR8 mutants). This approach alleviated the lack of in-depth miRNA annotations for some model organisms (Saçar *et al.*, 2013). We have shown using the comparison of wildtype and mutant datasets, in the cumulative plots, that miRCat2 generally performs better than all other tools tested, both overall and when confirming novel annotations. The tool also remains consistent in its predictions across all animal and plant data whilst the other tools tend to perform better only on some of the organisms: miRCat and miRDeep2 perform well in *H. sapiens* and *D. rerio*, while miRPlant performs well in *A. thaliana*.

miRCat2 is based on a new peak selection and feature-filtering algorithm i.e. it can only detect miRNAs with conservative secondary structures and miRNA-specific features. In animals, the pre-miRNAs have a well-defined structure with little fluctuations, making the detection of miRNAs easier. In plants, however, there is a higher degree of variability in miRNA hairpin length (Cuperus *et al.*, 2011) and hairpins can contain multiple loops and additional smaller hairpins (Chen, 2005; Xie *et al.*, 2015). These features make the plant miRNA detection challenging. Therefore, rule-based tools, such as miRCat2, miRCat, miRDeep2, miRPlant and miReap, may perform poorly on plant data, missing miRNAs with uncharacteristic features or allowing a large number of false positives. The results for plant data show that miReap performs poorly, displaying low sensitivity and specificity and also the poorest performance in the comparison with mutant datasets. This indicates high false positive and false negative rates and, although it performs better on animal data, miReap should probably not be used for plant miRNA prediction.

Another criterion that influences the outcome of miRCat2 is the read abundance of a miRNA locus: miRCat2 may miss miRNAs that are lowly expressed in the input samples due to the calculations used to test against a random uniform distribution, for the identification of peaks. Nevertheless the detection of low abundance miRNAs is a common issue for all miRNA prediction tools. This is not necessarily a disadvantage, as low read counts would suggest that the miRNA may not be expressed in that particular sample. In another sample where the miRNA is more highly expressed it is more likely that it would be predicted.

miRCat2 generates a score as a mean of ranking its predictions and performs well irrespective of a filtering based on this score. This suggests that the core algorithm is robust.

In terms of run time, miRCat2 compares favourably with miRDeep2, although miReap was faster. For example, on a *H. sapiens* dataset, containing approximately 34.5 million reads, miRCat2 generated the results in 3h50m, while miRDeep2 generated the results in 5h15m (all tests performed on a Linux server with CentOS 5.11 operating system, 144GB of memory and 2 Intel Xeon X5550 processors). In terms of memory usage, the amount allocated for one miRCat2 run is user-defined making it versatile to run on a wide range of specifications.

In conclusion, miRCat2 provides improved identification and characterization of new miRNAs over a range of organisms, that are not predicted by other tools. It should therefore contribute to a better, more in depth understanding of miRNAs, both in plants and animals.

FUNDING

This work was supported by Biotechnology and Biological Sciences Research Council (BBSRC) (V.M., M.S. and T.D., BB/L021269/1)

Conflict of interest statement.

None declared.

References

Ahmed,A., Ward,N.J., Moxon,S., Lopez-Gomollon,S., Viaut,C., Tomlinson,M.L., Patrushev,I., Gilchrist,M.J., Dalmay,T., Dotlic,D. *et al.* (2015) A Database of microRNA Expression Patterns in Xenopus laevis. *PloS one,* **10** (10), e0138313.

Ameres,S.L. and Zamore,P.D. (2013) Diversifying microRNA sequence and function. *Nature reviews Molecular cell biology,* **14** (8), 475–488.

An,J., Lai,J., Lehman,M.L. and Nelson,C.C. (2013) miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic acids research,* **41** (2), 727–737.

An,J., Lai,J., Sajjanhar,A., Lehman,M.L. and Nelson,C.C. (2014) miRPlant: an integrated tool for identification of plant miRNA from RNA sequencing data. *BMC bioinformatics,* **15** (1), 275.

Baker,M. (2010) Next-generation sequencing: adjusting to data overload. *nature methods,* **7** (7), 495–499.

Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets - update. *Nucleic acids research,* **41** (D1), D991–D995.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *cell,* **116** (2), 281–297.

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *cell,* **136** (2), 215–233.

Bazzini,A.A., Lee,M.T. and Giraldez,A.J. (2012) Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science,* **336** (6078), 233–237.

Bernstein,E., Caudy,A.A., Hammond,S.M. and Hannon,G.J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature,* **409** (6818), 363–366.

Bohnsack,M.T., Czaplinski,K. and GÖRLICH,D. (2004) Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. *Rna,* **10** (2), 185–191.

Bonnet,E., Wuyts,J., Rouzé,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** (17), 2911–2917.

Bosson,A.D., Zamudio,J.R. and Sharp,P.A. (2014) Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Molecular cell,* **56** (3), 347–359.

Cai,X., Hagedorn,C.H. and Cullen,B.R. (2004) Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna,* **10** (12), 1957–1966.

Chen,X. (2005) MicroRNA biogenesis and function in plants. *FEBS letters,* **579** (26), 5923–5931.

Cheng,A.M., Byrom,M.W., Shelton,J. and Ford,L.P. (2005) Antisense inhibition of human miRNAs and indications for an involvement of miRNA in cell growth and apoptosis. *Nucleic acids research,* **33** (4), 1290–1297.

Cifuentes,D., Xue,H., Taylor,D.W., Patnode,H., Mishima,Y., Cheloufi,S., Ma,E., Mane,S., Hannon,G.J., Lawson,N.D. *et al.* (2010) A novel miRNA processing pathway independent of Dicer requires Argonaute2 catalytic activity. *Science,* **328** (5986), 1694–1698.

Cuperus,J.T., Fahlgren,N. and Carrington,J.C. (2011) Evolution and functional diversification of MIRNA genes. *The Plant Cell,* **23** (2), 431–442.

Curtin,S.J., Michno,J.M., Campbell,B.W., Gil-Humanes,J., Mathioni,S.M., Hammond,R., Gutierrez-Gonzalez,J.J., Donohue,R.C., Kantar,M.B., Eamens,A.L. *et al.* (2016) MicroRNA maturation and microRNA target gene expression regulation are severely disrupted in soybean dicer-like1 double mutants. *G3: Genes\ Genomes\ Genetics,* **6** (2), 423–433.

Denli,A.M., Tops,B.B., Plasterk,R.H., Ketting,R.F. and Hannon,G.J. (2004) Processing of primary microRNAs by the Microprocessor complex. *Nature,* **432** (7014), 231–235.

Dillies,M.A., Rau,A., Aubert,J., Hennequet-Antier,C., Jeanmougin,M., Servant,N., Keime,C., Marot,G., Castel,D., Estelle,J. *et al.* (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in bioinformatics,* **14** (6), 671–683.

Djuranovic,S., Nahvi,A. and Green,R. (2012) miRNA-mediated gene silencing by translational repression followed by mRNA deadenylation and decay. *Science,* **336** (6078), 237–240.

Eamens,A.L., Smith,N.A., Curtin,S.J., Wang,M.B. and Waterhouse,P.M. (2009) The Arabidopsis thaliana double-stranded RNA binding protein DRB1 directs guide strand selection from microRNA duplexes. *Rna,* **15** (12), 2219–2235.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs - microRNAs with a role in cancer. *Nature Reviews Cancer,* **6** (4), 259–269.

Evers,M., Huttner,M., Dueck,A., Meister,G. and Engelmann,J.C. (2015) miRA: adaptable novel miRNA identification in plants using small RNA sequencing data. *BMC bioinformatics,* **16** (1), 1.

Fabian,M.R. and Sonenberg,N. (2012) The mechanics of miRNA-mediated gene silencing: a look under the hood of miRISC. *Nature structural & molecular biology,* **19** (6), 586–593.

Friedländer,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nature biotechnology,* **26** (4), 407–415.

Friedländer,M.R., Lizano,E., Houben,A.J., Bezdan,D., Báñez-Coronel,M., Kudla,G., Mateu-Huertas,E., Kagerbauer,B., González,J., Chen,K.C. *et al.* (2014) Evidence for the biogenesis of more than 1,000 novel human microRNAs. *Genome Biol,* **15** (4), R57.

Friedländer,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic acids research,* **40** (1), 37–52.

Garcia-Segura,L., Abreu-Goodger,C., Hernandez-Mendoza,A., Dinkova,T.D.D., Padilla-Noriega,L., Perez-Andrade,M.E. and Miranda-Rios,J. (2015) High-Throughput Profiling of Caenorhabditis elegans Starvation-Responsive microRNAs. *PloS one,* **10** (11), e0142262.

Gregory,R.I., Yan,K.p., Amuthan,G., Chendrimada,T., Doratotaj,B., Cooch,N. and Shiekhattar,R. (2004) The Microprocessor complex mediates the genesis of microRNAs. *Nature,* **432** (7014), 235–240.

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

Groenendyk,J., Fan,X., Peng,Z., Ilnytskyy,Y., Kurgan,L. and Michalak,M. (2014*a*) Genome-wide analysis of thapsigargin-induced microRNAs and their targets in NIH3T3 cells. *Genomics data,* **2**, 325–327.

Groenendyk,J., Hetz,C., Kurgan,L. and Michalak,M. (2014*b*) P125Endoplasmic reticulum stress responses to disrupted endoplasmic reticulum ca2+ homeostasis. *Cardiovascular research,* **103** (suppl 1), S22–S22.

Ha,M. and Kim,V.N. (2014) Regulation of microRNA biogenesis. *Nature reviews Molecular cell biology,* **15** (8), 509–524.

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 research,* **37** (suppl 2), W68–W76.

Han,J., Lee,Y., Yeom,K.H., Kim,Y.K., Jin,H. and Kim,V.N. (2004) The Drosha-DGCR8 complex in primary microRNA processing. *Genes & development,* **18** (24), 3016–3027.

Higashi,S., Fournier,C., Gautier,C., Gaspin,C. and Sagot,M.F. (2015) Mirinho: An efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing

data. *BMC bioinformatics,* **16** (1), 1.

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** (2), 232–243.

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

Iorio,M.V., Ferracin,M., Liu,C.G., Veronese,A., Spizzo,R., Sabbioni,S., Magri,E., Pedriali,M., Fabbri,M., Campiglio,M. *et al.* (2005) MicroRNA gene expression deregulation in human breast cancer. *Cancer research,* **65** (16), 7065–7070.

Jones-Rhoades,M.W., Bartel,D.P. and Bartel,B. (2006) MicroRNAs and their regulatory roles in plants. *Annu. Rev. Plant Biol.,* **57**, 19–53.

Kang,W. and Friedländer,M.R. (2015) Computational prediction of miRNA genes from small RNA sequencing data. *Frontiers in bioengineering and biotechnology,* **3**, 7.

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

Khvorova,A., Reynolds,A. and Jayasena,S.D. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell,* **115** (2), 209–216.

Kim,V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nature reviews Molecular cell biology,* **6** (5), 376–385.

Kim,Y.K., Kim,B. and Kim,V.N. (2016) Re-evaluation of the roles of DROSHA, Exportin 5, and DICER in microRNA biogenesis. *Proceedings of the National Academy of Sciences,* **113** (13), E1881–E1889.

Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic acids research,* **42** (D1), D68–D73.

Kravchik,M., Sunkar,R., Damodharan,S., Stav,R., Zohar,M., Isaacson,T. and Arazi,T. (2014) Global and local perturbation of the tomato microRNA pathway by a trans-activated DICER-LIKE 1 mutant. *Journal of experimental botany,* **65** (2), 725–739.

Kullback,S. and Leibler,R.A. (1951) On information and sufficiency. *The annals of mathematical statistics,* **22** (1), 79–86.

Kurihara,Y. and Watanabe,Y. (2004) Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proceedings of the National Academy of Sciences of the United States of America,* **101** (34), 12753–12758.

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** (5543), 858–862.

Lee,J., Kim,D.i., Park,J.H., Choi,I.Y. and Shin,C. (2013) miRAuto: An automated user-friendly MicroRNA prediction tool utilizing plant small RNA sequencing data. *Molecules and cells,* **35** (4), 342–347.

Lee,M., Choi,Y., Kim,K., Jin,H., Lim,J., Nguyen,T.A., Yang,J., Jeong,M., Giraldez,A.J., Yang,H. *et al.* (2014) Adenylation of maternally inherited microRNAs by Wispy. *Molecular cell,* **56** (5), 696–707.

Lee,Y., Ahn,C., Han,J., Choi,H., Kim,J., Yim,J., Lee,J., Provost,P., Rådmark,O., Kim,S. *et al.* (2003) The nuclear RNase III Drosha initiates microRNA processing. *Nature,* **425** (6956), 415–419.

Lee,Y., Kim,M., Han,J., Yeom,K.H., Lee,S., Baek,S.H. and Kim,V.N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *The EMBO journal,* **23** (20), 4051–4060.

Lei,J. and Sun,Y. (2014) mir-prefer: an accurate, fast and easy-to-use plant mirna prediction tool using small rna-seq data. *Bioinformatics,* , btu380.

Leinonen,R., Sugawara,H. and Shumway,M. (2010) The sequence read archive. *Nucleic acids research,* , gkq1019.

Li,Y., Zhang,Z., Liu,F., Vongsangnak,W., Jing,Q. and Shen,B. (2012) Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis. *Nucleic acids research,* , gks043.

Lopez-Gomollon,S., Mohorianu,I., Szittya,G., Moulton,V. and Dalmay,T. (2012) Diverse correlation patterns between microRNAs and their targets during tomato fruit development indicates different modes of microRNA actions. *Planta,* **236** (6), 1875–1887.

Lorenz,R., Bernhart,S.H., Zu Siederdissen,C.H., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms for Molecular Biology,* **6** (1), 1.

Lu,M., Zhang,Q., Deng,M., Miao,J., Guo,Y., Gao,W. and Cui,Q. (2008) An analysis of human microRNA and disease associations. *PloS one,* **3** (10), e3420.

Lund,E., Güttinger,S., Calado,A., Dahlberg,J.E. and Kutay,U. (2004) Nuclear export of microRNA precursors. *Science,* **303** (5654), 95–98.

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** (18), 2226–2234.

McCormick,K.P., Willmann,M.R. and Meyers,B.C. (2011) Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments. *Silence,* **2** (1), 1.

Meng,X., Yang,S., Zhang,Y., Wang,X., Goodfellow,R.X., Jia,Y., Thiel,K.W., Reyes,H.D., Yang,B. and Leslie,K.K. (2015) Genetic deficiency of Mtdh gene in mice causes male infertility via impaired spermatogenesis and alterations in the expression of small non-coding RNAs. *Journal of Biological Chemistry,* **290** (19), 11853–11864.

Meyers,B.C., Axtell,M.J., Bartel,B., Bartel,D.P., Baulcombe,D., Bowman,J.L., Cao,X., Carrington,J.C., Chen,X., Green,P.J. *et al.* (2008) Criteria for annotation of plant MicroRNAs. *The Plant Cell,* **20** (12), 3186–3190.

Modzelewski,A.J., Hilz,S., Crate,E.A., Schweidenback,C.T., Fogarty,E.A., Grenier,J.K., Freire,R., Cohen,P.E. and Grimson,A. (2015) Dgcr8 and Dicer are essential for sex chromosome integrity during meiosis in males. *J Cell Sci,* **128** (12), 2314–2327.

Mohorianu,I., Schwach,F., Jing,R., Lopez-Gomollon,S., Moxon,S., Szittya,G., Sorefan,K., Moulton,V. and Dalmay,T. (2011) Profiling of short RNAs during fleshy fruit development reveals stage-specific sRNAome expression patterns. *The Plant Journal,* **67** (2), 232–246.

Mohorianu,I., Stocks,M.B., Wood,J., Dalmay,T. and Moulton,V. (2013) CoLIde: a bioinformatics tool for CO-expression based small RNA Loci Identification using high-throughput sequencing data. *RNA biology,* **10** (7), 1221–1230.

Mortazavi,A., Williams,B.A., McCue,K., Schaeffer,L. and Wold,B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods,* **5** (7), 621–628.

Moxon,S., Schwach,F., Dalmay,T., MacLean,D., Studholme,D.J. and Moulton,V. (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics,* **24** (19), 2252–2253.

Noh,H., Park,C., Park,S., Lee,Y.S., Cho,S.Y. and Seo,H. (2014) Prediction of miRNA-mRNA associations in Alzheimer's disease mice using network topology. *BMC genomics,* **15** (1), 1.

Papp,I., Mette,M.F., Aufsatz,W., Daxinger,L., Schauer,S.E., Ray,A., Van Der Winden,J., Matzke,M. and Matzke,A.J. (2003) Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant physiology,* **132** (3), 1382–1390.

Park,W., Li,J., Song,R., Messing,J. and Chen,X. (2002) CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Current Biology,* **12** (17), 1484–1495.

Pérez-Quintero,Á.L., Neme,R., Zapata,A. and López,C. (2010) Plant microRNAs and their role in defense against viruses: a bioinformatics approach. *BMC plant biology,* **10** (1), 1.

Prüfer,K., Stenzel,U., Dannemann,M., Green,R.E., Lachmann,M. and Kelso,J. (2008) PatMaN: rapid alignment of short sequences to large databases. *Bioinformatics,* **24** (13), 1530–1531.

Reinhart,B.J., Weinstein,E.G., Rhoades,M.W., Bartel,B. and Bartel,D.P. (2002) MicroRNAs in plants. *Genes & development,* **16** (13), 1616–1626.

Saçar,M.D., Hamzeiy,H. and Allmer,J. (2013) Can MiRBase provide positive data for machine learning for the detection of MiRNA hairpins. *Journal of integrative bioinformatics,* **10** (2), 215.

Shin,C., Nam,J.W., Farh,K.K.H., Chiang,H.R., Shkumatava,A. and Bartel,D.P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular cell,* **38** (6), 789–802.

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 research,* **20** (9), 1207–1218.

Stocks,M.B., Moxon,S., Mapleson,D., Woolfenden,H.C., Mohorianu,I., Folkes,L., Schwach,F., Dalmay,T. and Moulton,V. (2012) The UEA sRNA workbench: a suite of tools for analysing and visualizing next generation sequencing microRNA and small RNA datasets. *Bioinformatics,* **28** (15), 2059–2061.

Surridge,A.K., Lopez-Gomollon,S., Moxon,S., Maroja,L.S., Rathjen,T., Nadeau,N.J., Dalmay,T. and Jiggins,C.D. (2011) Characterisation and expression of microRNAs in developing wings of the neotropical butterfly Heliconius melpomene. *BMC genomics,* **12** (1), 1.

Tucker,T., Marra,M. and Friedman,J.M. (2009) Massively parallel sequencing: the next big thing in genetic medicine. *The American Journal of Human Genetics,* **85** (2), 142–154.

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** (1), 1.

Wang,H., Zhang,X., Liu,J., Kiba,T., Woo,J., Ojo,T., Hafner,M., Tuschl,T., Chua,N.H. and Wang,X.J. (2011) Deep sequencing of small RNAs specifically associated with Arabidopsis AGO1 and AGO4 uncovers new AGO functions. *THe plant journal,* **67** (2), 292–304.

Williamson,V., Kim,A., Xie,B., McMichael,G.O., Gao,Y. and Vladimirov,V. (2012) Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Briefings in bioinformatics,* , bbs010.

Wu,L., Zhang,Q., Zhou,H., Ni,F., Wu,X. and Qi,Y. (2009) Rice MicroRNA effector complexes and targets. *The Plant Cell,* **21** (11), 3421–3435.

Xie,M., Zhang,S. and Yu,B. (2015) microRNA biogenesis, degradation and activity in plants. *Cellular and Molecular Life Sciences,* **72** (1), 87–99.

Xie,Z., Johansen,L.K., Gustafson,A.M., Kasschau,K.D., Lellis,A.D., Zilberman,D., Jacobsen,S.E. and Carrington,J.C. (2004) Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol,* **2** (5), e104.

Xuan,P., Guo,M., Huang,Y., Li,W. and Huang,Y. (2011) MaturePred: efficient identification of microRNAs within novel plant pre-miRNAs. *PloS one,* **6** (11), e27422.

Yang,X. and Li,L. (2011) miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics,* **27** (18), 2614–2615.

Yi,R., Qin,Y., Macara,I.G. and Cullen,B.R. (2003) Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & development,* **17** (24), 3011–3016.

Yu,B., Yang,Z., Li,J., Minakhina,S., Yang,M., Padgett,R.W., Steward,R. and Chen,X. (2005) Methylation as a crucial step in plant microRNA biogenesis. *Science,* **307** (5711), 932–935.

Zeng,Y., Yi,R. and Cullen,B.R. (2005) Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *The EMBO journal,* **24** (1), 138–148.