OXFORD

Briefings in Bioinformatics, 20(3), 2019, 918–930

doi: 10.1093/bib/bbx148
Advance Access Publication Date: 6 November 2017
Paper

A survey of software tools for microRNA discovery and characterization using RNA-seq

Michele Bortolomeazzi, Enrico Gaffo and Stefania Bortoluzzi

Corresponding authors: Stefania Bortoluzzi, Department of Molecular Medicine, University of Padova, Via G. Colombo 3, 35131, Padova, Italy. E-mail: stefania.bortoluzzi@unipd.it; Enrico Gaffo, Department of Molecular Medicine, University of Padova, Via G. Colombo 3, 35131, Padova, Italy. E-mail: enrico.gaffo@unipd.it

Abstract

Since the small RNA-sequencing (sRNA-seq) technology became available, it allowed the discovery of thousands new microRNAs (miRNAs) in humans and many other species, providing new data on these small RNAs (sRNAs) of high biological and translational relevance. MiRNA discovery has not yet reached saturation, even in the most studied model organisms, and many researchers are using sRNA-seq in studies with different aims in biomedicine, fundamental research and in applied animal sciences. We review several miRNA discovery and characterization software tools that implement different strategies, providing a useful guide for researchers to select the programs best suiting their study objectives and data. After a brief introduction on miRNA biogenesis, function and characteristics, useful to understand the biological background considered by the algorithms, we survey the current state of miRNA discovery bioinformatics discussing 26 different sRNA-seq-based miRNA prediction software and toolkits released in the past 6 years, including 15 methods specific for miRNA prediction and 11 more general-purpose software suites for sRNA-seq data analysis. We highlight the main features of mature miRNAs and miRNA precursors considered by the methods categorizing them according to prediction strategy and implementation. In addition, we describe a typical miRNA prediction and analysis workflow by delineating the objectives, potentialities and main steps of sRNA-seq data analysis projects, from preparatory data processing to miRNA prediction, quantification and diverse downstream analyses. Finally, we outline the caveats affecting sRNA-seq-based prediction tools, and we indicate the possibilities offered by data set pooling and by integration with other types of high-throughput sequencing data.

Key words: microRNA; RNA-seq; miRNA discovery; bioinformatics tools; moRNA; IsomiRs

Introduction

MicroRNAs (miRNAs) are ∼22 nt single-stranded small noncoding RNAs (sncRNAs) that negatively regulate mRNA stability and/or translation by recruiting different combinations of effector proteins on target transcripts [1, 2]. The effects of miRNA expression can be pervasive, as one single miRNA can target mRNAs of hundreds of different genes. MiRNAs are involved in most physiological processes, and play important roles in cell-fate determination and development [3–5]. Most known miRNAs are evolutionarily conserved and their number largely correlates with organism complexity [6].

In humans, deregulation or alteration of miRNAs plays functional roles in numerous pathological conditions [7], including cancer [8–10]. Interestingly, miRNA profiles can predict relevant tumor subtypes, treatment response and outcome [11]. For these characteristics, miRNAs have been thoroughly investigated both as therapeutic targets [12, 13] and noninvasive biomarkers [14]. Promising oncology works studied the use of circulating miRNAs in patient plasma as biomarkers [15], and on small RNAs (sRNAs) carried by vesicles as possible disease modulators or preconditioning elements in the process of metastasis formation [16]. Beyond the biomedical field, also many animal and food genomics research projects are studying miRNAs [17–19].

As reported in the latest release of miRBase (v.21, June 2014), the most popular reference database of miRNA sequences, almost 36 000 miRNAs have been detected in 223 species: from

Michele Bortolomeazzi recently graduated in Molecular Biology at the University of Padova with a thesis on miRNA discovery.
Enrico Gaffo is a postdoc with background in Computer Science and experience in NGS transcriptomics.
Stefania Bortoluzzi is a professor of Applied Biology at the Molecular Medicine Department of the University of Padova, where she leads a computational genomics group.

Submitted: 14 July 2017; **Received (in revised form):** 12 October 2017

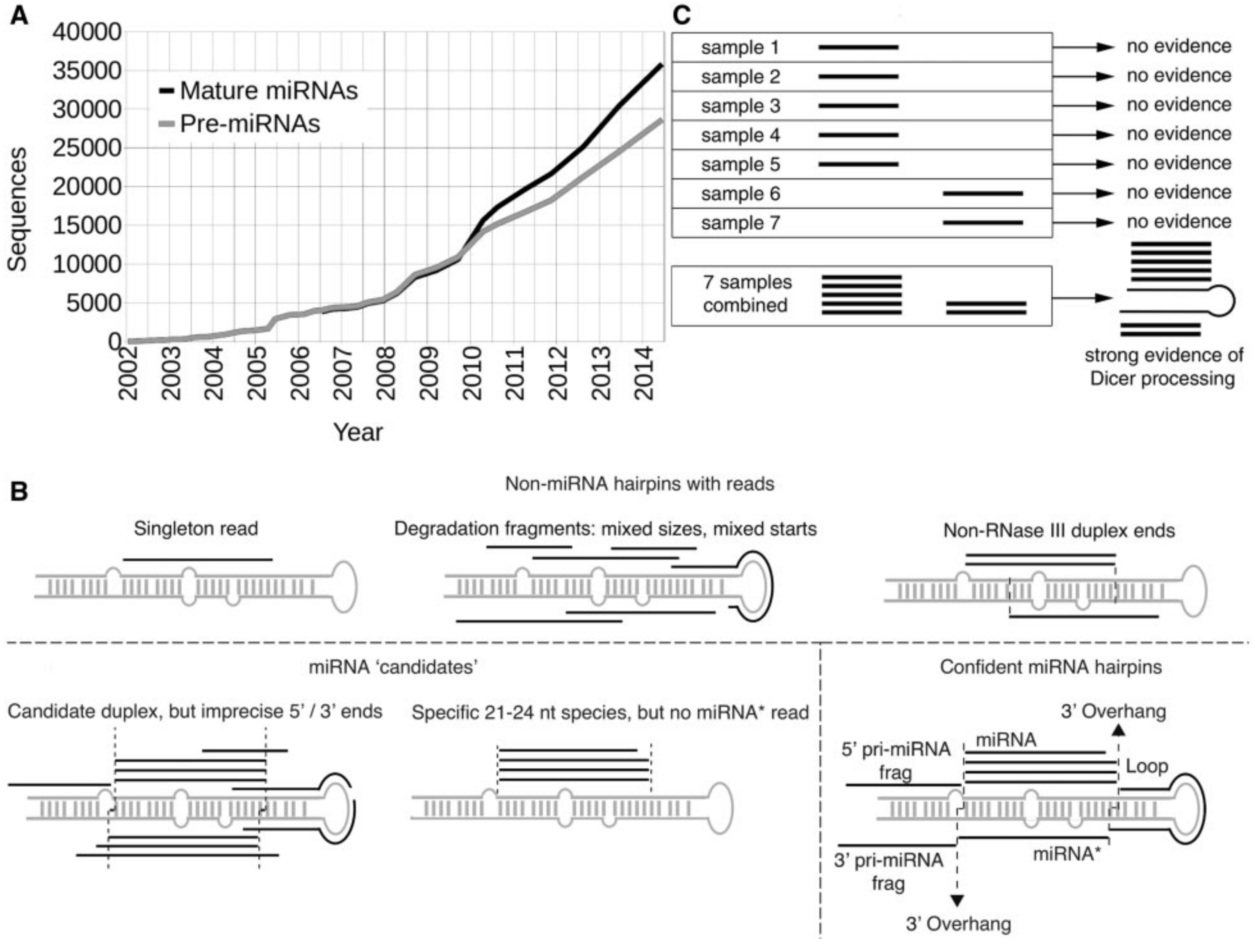

Figure 1. miRNA discovery from sRNA-seq experiments. (**A**) Number of known mature miRNAs (black) and pre-miRNAs (gray) present in miRBase by year, since 2002. (**B**) Examples of the three main cases of putative precursors' read signatures: the lack of reads corresponding to any of the two mature miRNAs corresponds to read signatures on on-miRNA hairpins; hairpin with read signatures typical of a true precursor, with precise 3′ overhangs and reads mapping to definite regions of the hairpin (i.e. aligning to the miRNA, the miRNA* or the loop without overlaps between different products) indicates high confidence miRNAs; an intermediate case in represented by *bona fide* pre-miRNAs with read signatures displaying high 5′ heterogeneity or the absence of miRNA* reads, which cannot be annotated with high confidence, and can be confirmed using additional evidence, such as read depletion in miRNA biogenesis mutants or enrichment in Ago-IP experiments (adapted from [21]). (**C**) Drawing illustrating how data set pooling can increase the sensitivity of miRNA prediction analyses, by allowing the generation of highly informative read signatures even for lowly expressed pre-miRNAs (adapted from [22]) .

algae and slime molds to higher eukaryotes and viruses. However, miRNA annotation is clearly incomplete for less studied metazoans: only 88 miRNA loci are known for *Pan paniscus*, whereas a number close to that in the human genome (over 1800) is expected. In addition, over 10 000 entries were added to miRBase in the 3-year interval between the current release (v.21, June 2014) and version 18 (v.18, November 2011) [20], including >4000 entries from the previous (v.20, June 2013) and the current release. Thus, the number of expected miRNAs being higher than current available annotations and the high deposition rate of novel miRNA annotations in recent years (Figure 1A) suggest that many miRNAs have not yet been discovered in model species and humans.

Bioinformatic prediction from small RNA-sequencing (sRNA-seq) data turned out to be a powerful approach for extensive miRNA detection and study. sRNA-seq data can be generated in the frame of different studies. According to the investigated organism and project aims, a rigorous and effective experimental design should consider optimal replicate number, proper controls, RNA extraction and library preparation protocol, spike-ins, platform, depth and cost of the sequencing. These aspects are fundamental for high-quality data production and were surveyed in other works [23–25]. Here, we focus on the tools for the bioinformatics analysis that follows the sRNA-seq data production.

Several programs for miRNA identification and quantification from sRNA-seq data are currently available [26, 27], and many of them are accessible even to laboratories with little bioinformatic expertise and computational resources. The currently available reviews on sRNA-seq data analysis focus specifically on miRNA discovery [28–30] or offer a wide view of the subject by covering also other topics such as miRNA target prediction [31, 32]. However, these studies include approaches now outdated, offering little guidance to prospective users. To integrate and complement these reviews, here we provide a summary of the state-of-the-art sRNA-seq-based software for the prediction of animal miRNAs, considering also its application to the identification of noncanonical miRNAs and miRNA isoforms. We briefly recall the main aspects of miRNA biogenesis and of non-next generation sequencing (NGS)-based discovery methods. Next, we introduce sRNA-seq-based discovery tools grouped by prediction strategy. We describe a typical miRNA prediction and analysis workflow and indicate the main toolkits implementing it. Finally, we discuss the challenges posed by the validation of hundreds of predicted miRNA sequences. We conclude with critical considerations on the illustrated methods that can be useful to choose the best one according to the study.

MiRNA biogenesis and function

To better understand the biological bases of the algorithms discussed below, we recapitulate the main aspects of miRNA biogenesis, functions and characteristics.

In the canonical pathway [33], miRNA genes are transcribed into long capped and polyadenylated transcripts, the pri-miRNAs, which are then cleaved by the nuclear type III endonuclease Drosha into 65–75 nt single-stranded hairpins called miRNA precursors, the pre-miRNAs. In addition, miRNAs can be generated through noncanonical mechanisms [34, 35]. For instance, pre-miRNAs can derive from spliced intronic sequences (miRtrons) [36] or from the processing of housekeeping noncoding RNAs (ncRNAs), such as small nucleolar RNAs (snoRNAs) [37, 38] and transfer RNAs (tRNAs) [39, 40]. Pre-miRNAs are transported in the cytosol where they are cleaved by Dicer, resulting in ∼22 nt double-stranded molecules. Afterward, this miRNA duplex is unwinded and one of the two strands, the mature miRNA, is loaded onto an Argonaute protein (usually AGO2) of the RNA-induced silencing complex (RISC). The selection of which strand of the duplex will form the mature miRNA is an actively regulated process that can vary across different tissues and conditions [33]. MiRNAs were formerly referred to as major miRNA (the one that is loaded onto the RISC) and miRNA* (the one that undergoes degradation), but because of evidence of functional miRNAs*, nowadays it is preferred to name them miR-5p or miR-3p, according to the originating pre-miRNA strand.

Mature miRNAs guide the RISC to target mRNAs by partial sequence complementarity. The canonical targeting involves particularly the miRNA nucleotides 2–7 or 2–8, the seed region and the mRNA 3′-untranslated region (UTR), even though different pairing types were observed as well [41]. MiRNA-mRNA interactions result either in the inhibition of translation (through still unclear mechanisms) or in the degradation of the mRNA, generally by promoting deadenylation, decapping and 5′-to-3′ decay [42–44]. In addition, miRNAs binding outside 3′-UTRs can exert noncanonical functions: for instance, miRNA binding to the 5′ region of a mRNA coding sequence can regulate alternative translation [45], and binding sites in transcript coding sequence can mediate a reduction of the target mRNA abundance [46].

Postprocessing nucleotide addition, removal and editing were observed in most miRNA sequences, so that miRNA expression can be viewed as a mixture of miRNA isoforms (isomiRs). The sequence variations involve mainly the 3′ end and can affect target specificity, stability and AGO-loading efficiency [47, 48], thus impacting on the miRNA function. IsomiR composition varies among tissue and conditions, and in several cases, the sequence in official annotation is not the most expressed form [49–51]. IsomiRs identified from sRNA-seq data by computational methods were experimentally validated in many species [49, 52–54].

Non-NGS-based miRNA discovery approaches

Early methods for miRNA discovery were laborious and of limited discovery power, as they relied on low-throughput experimental procedures that required isolating the molecules on high-resolution gels, cloning and Sanger sequencing. To support the findings, the sequences were further assessed by computing the probability of folding into hairpin-like structures, and by requiring no other ncRNA loci to overlap [55, 56].

Next, bioinformatic methods were developed to predict miRNA loci by scanning the genomic sequence for hairpin-forming sequences [57]. These genome-based approaches grounded either on evolutionary conservation, or on similarity with known miRNA loci.

Refining predictions by considering evolutionary conservation is a common strategy in computational biology, as functional elements are under selection. Pre-miRNAs present a particular conservation pattern: the strand regions are more conserved than the loop, and the miRNA seed nucleotides are the most conserved. In general, a miRNA is considered conserved when its seed sequence is identical to those of its orthologs in other species. Assuming these characteristics, the MiRSeeker [58] and miRScan [59] miRNA prediction methods select the loci with sequences that can fold into a hairpin-like structure, and that have miRNA homologs in other species.

Similarity-based methods identify loci with sequence and RNA secondary structure similar to known miRNAs by considering several features including: sequence entropy and composition, folding energy and hairpin-like structure and number of paired bases in the predicted miRNA duplex. Most of these programs implement machine learning approaches based on different classifiers, such as hierarchical hidden Markov models (HHMMiR [60]) and support vector machines (SVMs) (mirnaDetect [61]). Other software use statistical analysis, like the scoring-based miRalign [62].

Both conservation and similarity-based approaches require previous information and are inherently biased toward the discovery of miRNAs homologous and/or similar to already known loci.

sRNA-seq-based miRNA prediction

Genome-based methods are subject to high rates of false-positive predictions, as only hundreds of the millions of loci whose transcripts could fold into hairpin structures are transcribed and processed by the miRNA biogenesis pathways [57].

Transcriptome sequencing resulted to be informative to correct this weakness: several programs used sRNA-seq reads to detect the candidate pre-miRNA genomic sequences to be evaluated with genome-based methods. Examples are the similarity-based miRD [63] and miR-BAG [64], both using machine learning strategies, the conservation-based MIRPIPE [65] and the rule-based (see below) approach MIRINHO [66]. These methods exploit sRNA-seq information only partially, as read alignments are used only for the excision of putative precursors, and pre-miRNA evaluation relies completely on the DNA sequence.

At present, prediction from sRNA-seq data is the most commonly used approach for miRNA discovery, and most of the known miRNA loci and mature miRNA sequences have been discovered with this technique [20, 67] (Figure 1A). In the past 6 years, several bioinformatic miRNA prediction tools have been developed and applied both in small studies and in large projects with tens or hundreds of data sets [68–70].

Features considered for sRNA-seq-based miRNA discovery

Algorithms for the prediction of novel animal miRNAs from sRNA-seq data are based on the evaluation of the read signature associated with each putative precursor, i.e. the distribution of reads on the precursor sequence, which allows to determine if the reads originated from a putative precursor are compatible with the canonical miRNA biogenesis pathway (Figure 1B). Expectedly, all the reads derived from a *bona fide* miRNA hairpin should correspond to either the mature miRNA, the miRNA* or the loop sequences. The predicted miRNA and miRNA* sequences should also form a duplex with 2 nt overhanging at the 3′ of both ends, as from the typical Dicer processing. Although some variability was observed [20], mature miRNAs' 5′ ends display in general low heterogeneity because of their origin from the RNase III-based endonucleolytic cropping of the precursors by Drosha and Dicer. Additionally, homogeneous 5′ ends are required for seed-based miRNA-targeting, as any variation at the 5′ end would impact target recognition specificity. Thus, miRNA prediction software requires that the reads corresponding to mature miRNAs have highly consistent 5′ ends. On the contrary, prediction algorithms generally tolerate 3′ end variability, as the 3′ ends of mature miRNAs are often heterogeneous in *bona fide* hairpins [47] because of alternative processing and/or successive nucleotide additions [47].

Reads derived from the antisense strand of a putative pre-miRNA were considered by early prediction methods, for instance (f.i.) the first version of miRDeep [71], to indicate false predictions, as antisense transcripts are more probably associated with other types of sncRNAs such as endo-siRNAs. However, more recent methods, such as miRDeep2 [72], do not penalize for antisense reads because several miRNAs are antisense to transcribed loci, and there is evidence of miRNA precursors pairs transcribed from opposite strands of the same locus, such as dme-mir-iab-4 and dme-mir-iab-8 in *Drosophila* [72]. Moreover, endo-siRNAs and piRNAs can be easily distinguished, as they are generally shifted by several bases from the corresponding locus on the opposite strand [73–75], while antisense miRNAs mostly present exact overlaps.

In addition to the features derived from pre-miRNA read signatures, described in the previous paragraph, most software also make use of structural information about the predicted pre-miRNAs and the miRNA duplex. The most probable RNA secondary structure of putative precursors is predicted, and the structural compatibility with canonical pre-miRNA hairpins is evaluated using: the minimum free energy (MFE) of the structure, the number of pairing bases between the miRNA and the sister miRNA sequence and the absence of branching or bulges outside the loop [28]. This evaluation can require the prediction of the precursor's RNA secondary structures, often performed with RNAfold [76] (Table 1), and the realignment of the reads onto a subset of putative pre-miRNA sequences.

MiRNA prediction software tools

Below, we discuss 15 currently available software tools for the prediction of novel animal miRNAs from sRNA-seq data (Table 1), implemented either as Web-based services or as programs than can be run locally, grouping them in three categories: read signature evaluation and read profile clustering, both relying on sRNA-seq read alignment to a reference genome, and miRNA duplex evaluation methods.

Evaluation of read signatures

Approaches evaluating read signatures first generate a set of putative precursors through the excision of ∼110 nt genomic sequences surrounding the highest local stack of aligned reads in a small sliding window, which is scanned along each strand of every chromosome. Afterward, the RNA secondary structure and MFE of every excised precursor are calculated together with several parameters describing both the read signature and the structures of the hairpin and the duplex. The most commonly used parameters are the number of matching nucleotides in the putative miRNA duplex, the length of putative miRNA duplex overhangs and the 5′ end entropy of miRNA reads. Finally, the excised sequences are evaluated as putative pre-miRNAs in different ways, according to one of the two prediction strategies described in the following paragraphs.

Rule-based prediction algorithms evaluate the parameters of each candidate precursor against reference values calculated from known pre-miRNAs and random sequences. For each precursor, these evaluations are finally summarized in a score, which quantifies the probability of being a true miRNA hairpin. The methods commonly used in novel miRNA prediction studies grounded on rule-based evaluation of read signatures include MIReNA [83], miRSeqNovel [81], miRdentify [82], miRDeep* [84] and miRDeep2 [72].

Machine learning-based methods require a training step on a set of read signature and structural features, calculated from known pre-miRNAs (positive set), and random hairpin-forming sequences (negative set), to generate a classification model that will be finally applied to the set of putative precursors. The selection of training data can have large impact on the quality of a classifier, in particular when it is used on data from species different from the one for which it was trained. However, for most species, there are not enough annotations to generate an adequate training set. Therefore, annotations have often to be pooled within a broad taxon, despite the resulting decline in sensitivity and specificity [93]. Examples of machine learning-based miRNA prediction software using random forest classifiers are CoRAL [80, 94] and miRanalyzer [79].

Read profile clustering

Blocks of reads mapping close to each other on the same genomic strand define read alignment profiles, or read profiles. The profiles are compared considering both the relative position and the length of the reads, to be clustered by similarity. Most software using the read profile clustering approach rely on the blockbuster tool [95] to collect consecutive reads with close start and end positions into sharp blocks. Then, profile clusters are annotated according to the known ncRNAs they overlap. NcRNA classification tools that analyze read blocks obtained with blockbuster include deepBlockAlign [87], BlockClust [86], DARIO [91] and miRDBA [92]. In contrast, FlaiMapper [88] uses a custom read-grouping algorithm.

Differently from methods based on evaluation of read signatures, approaches using read profiles clustering do not rely on the prediction and evaluation of putative hairpin secondary structures. Notably, apart from miRNAs, they can also identify from RNA sequencing (RNA-seq) data several other classes of small ncRNAs such as ribosomal RNAs, tRNAs and snoRNAs.

MiRNA duplex evaluation

Methods evaluating miRNA duplexes assemble the sRNA-seq reads into contigs, select those 10–30 nt long and match them into all possible pairs to generate putative miRNA duplexes. Afterward, a set of features including length, number of unpaired bases and overhangs is calculated to feed machine learning algorithms that select the most probably true miRNA duplexes. Examples of programs applying this method are MirPlex [77], which uses an SVM classifier, and miReader [78], which relies on a multi-boosting algorithm with Best-First Tree as base classifier.

Table 1. Software for miRNA prediction from sRNA-seq data

Name	Citation	URL	Latest release	Implementation	User interface	Strategy	Preprocessing	Mapping	Structure prediction	Other ncRNA	Differential expression	IsomiRs	Validation
miRPlex	[77]	www.uea.ac.uk/computing/mirplex	0.1, 2012	SA	CLI	MDE	No	Not used	Not used	No	No	No	–
miReader	[78]	scbb.ihbt.res.in/2810-12/miReader.php	2.0, 2016	SA	CLI, GUI	MDE	No	Not used	Not used	No	No	Yes	–
miRanalyzer	[79]	bioinfo5.ugr.es/miRanalyzer/miRanal yzer.php	0.3, 2012	SA, WS	CLI, GUI	ML	Yes	Bowtie	RNAfold	Filtered	DESeq	Yes	Dicer knockdown [72]
CoRAL	[80]	wanglab.pcbi.upenn.edu/coral/	1.1.1, 2013	SA, WS	CLI	ML	No	Bowtie	RNAfold	Classified	No	No	–
miRSeqnovel	[81]	sourceforge.net/projects/mir seq/files/	1.3, 2011	RP	CLI	RB	No	No	RNAfold	Filtered	edgeR, DESeq, custom	Yes	qRT-PCR [81]
miRDentify	[82]	www.ncrnalab.dk/#mirdentify/mirdentify.php	1.00, 2014	SA	CLI	RB	Yes	Bowtie	MultiRNAfold	Filtered	No	No	Northern blot [82]
MIReNA	[83]	www.lcqb.upmc.fr/mirena/index.html	2.0, 2010	SA	CLI, GUI	RB	No	megaBlast	RNAfold	Filtered	No	No	Dicer knockdown [72]
miRDeep2	[72]	www.mdc-berlin.de/8551903/en/	2.0.0.8, 2016	SA	CLI	RB	Yes	Bowtie	RNAfold	Filtered	No	No	Dicer knockdown [72]
miRDeep*	[84]	www.australianprostatecentre.org/research/software/mirdeep-star	37, 2016	SA	CLI, GUI	RB	Yes	Custom	Custom	Filtered	No	No	Dicer knockdown, qRT-PCR [84]
sRNAbench	[85]	bioinfo2.ugr.es: 8080/ceUGR/srnabench/	10/14 07/10/2014	SA	CLI	RB, ML	Yes	Bowtie	RNAfold	Filtered	edgeR	Yes	–
BlockClust	[86]	toolshed.g2.bx.psu.edu/reposi tory?repository_id=f3d4583fe94434e9	0.1, 2015	GW	CLI	RPC	No	No	Not used	Classified	No	No	–
deepBlockAlign	[87]	rth.dk/resources/dba/down load.php	1.3.1 2014	SA	CLI	RPC	No	No	Not used	Classified	No	No	–
FlaiMapper	[88]	github.com/yhoogstrate/flaimapper	2.3.4, 2016	SA	CLI	RPC	No	No	Not used	Classified	No	No	qRT-PCR [89, 90]
DARIO	[91]	dario.bioinf.uni-leipzig.de/index.py	2011	WS	GUI	RPC	No	No	Not used	Classified	No	No	–
miRDBA	[92]	rth.dk/resources/mirdba/	1.0, 2013	WS	GUI	RPC	No	No	Not used	Classified	No	No	–

Note: For each miRNA prediction tool, the table provides references, main features and validation studies, if available. Latest release: number and release date of the last version of the program. Implementation: SA, stand-alone; WS, Web server; RP, R Package; GW, Galaxy Workflow. User Interface: GUI, graphical user interface; CLI, command-line interface. Strategy: RB, rule-based, ML, machine learning; RPC, read profile clustering; MDE, miRNA duplex evaluation. Preprocessing: whether the program performs any read preprocessing steps on the input data. Mapping: short-read aligner used, in case the prediction software performs reads to genome mapping. Structure prediction: whether the application computes or requires the RNA secondary structure of putative precursor sequence (program used for the prediction where applicable). Other ncRNA: whether the program filters out or keeps and classifies reads belonging to non-miRNA ncRNA classes. Differential expression: whether the analysis is performed and with which tool, where applicable. IsomiRs: whether the software identifies and classifies isomiRs. Validation: citation of studies providing validation of sRNAs predicted with the software, and experimental methods used.

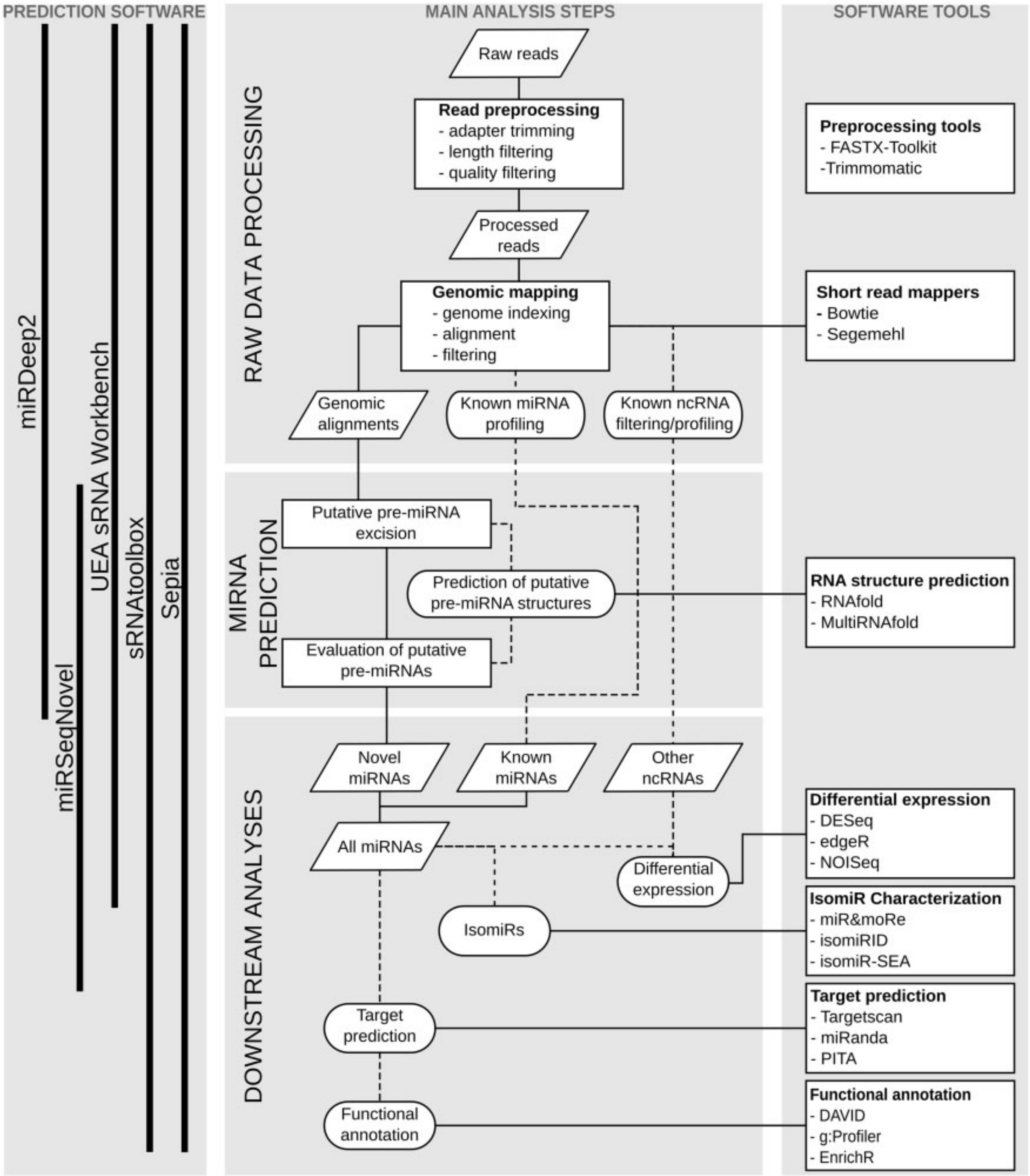

Figure 2. sRNA-seq-based miRNA prediction workflow and software tools. The central column contains a flowchart illustrating the steps of each of the three main phases of a typical miRNA discovery and characterization workflow (rectangles and continuous lines: analysis steps, parallelograms: data, rounded rectangles and dashed lines: optional steps). The left panel shows examples of miRNA prediction software and toolkits, with vertical lines indicating the workflow parts performed by each program. Tools commonly used to carry out specific tasks in each phase of the analysis' workflow are reported in the right panel.

MiRNA duplex evaluation programs are the only prediction methods that can be applied to species for which a reference genomic sequence is not yet available. However, this approach can identify only the mature miRNA and miRNA* products, and does not provide any information on the pre-miRNA hairpins. Another important limitation is that these programs require the presence of both miRNA and miRNA* reads: this could prevent the detection of miRNA duplexes, as often only one of the two mature miRNAs is present at a detectable level in a cell type or condition. Finally, the reliance on machine learning algorithms requiring positive training sets of miRNAs from closely related species limits the usefulness of these methods for understudied and nonmodel species.

MiRNA prediction and analysis workflow

Conceptually, the general sRNA-seq analysis workflow can be divided in three stages (Figure 2): preparatory data processing, miRNA prediction and downstream analyses.

Preparatory data processing

The first step preprocesses raw reads (usually in FASTQ format) by trimming sequencing adapters and selecting the trimmed reads according to both length (generally between 18 and 25 nt) and sequence quality. This step is performed automatically by

Table 2. Toolkits for sRNA-seq data analysis allowing miRNA prediction and characterization

Name	Citation	URL	Latest release	Implementation	User interface	miRNA prediction	Known ncRNAs	Target prediction	Functional analysis	Differential expression	IsomiRs	Other data
UEA sRNA Workbench	[98]	srna-workbench.cmp.uea.ac.uk/	4.3, 2016	SA	GUI, CLI	miRCat 2.0	Filtered	No	No	Custom	Yes	PARE-seq (target prediction)
wapRNA	[99]	waprna.big.ac.cn/	2012	SA, WS	GUI, CLI	miRDeep	Quantified	miRanda, RNAHybrid	Yes	DEGSeq	No	RNA-seq (mRNA)
OmiRas	[100]	tools.genxpro.net/omiras/	2013	WS	GUI	miRDeep	Quantified	Known miRNA only	known miRNA only	DESeq	No	No
eRNA	[101]	bioinformaticstools.mayo.edu/research/cap-mirseq/	1.01, 2014	SA	GUI, CLI	miRDeep2	Filtered	No	No	DESeq	Yes	RNA-seq (mRNA)
CAP-miRSeq	[102]	bioinformaticstools.mayo.edu/research/cap-mirseq/	2014	SA, VM	CLI	miRDeep2	Quantified	No	No	edgeR	Variants with GATK	No
miARMA-Seq	[103]	miarmaseq.cbbio.es/	1.6, 2016	SA, VM, DI	CLI	miRDeep2	Quantified	miRGate	Yes	edgeR, NOISeq	No	RNA-seq (circular RNAs, mRNAs)
Oasis2.0	[104]	oasis.dzne.de/	2.0, 2016	WS	GUI, API	miRDeep2	Quantified	miRanda	Yes	DESeq2	No	No
iMiR	[105]	www.labmedmolge.unisa.it/inglese/research/imir	2015	SA	GUI	miRDeep2, miRanalyser	Filtered	TargetScan, miRanda	No	DESeq	Yes	No
Sepia	[106]	anduril.org/sepia/index.html	20/05/2016	SA, DI	CLI	miRDeep2, miRanalyser	Custom	Custom	Yes	Custom	Custom	RNA-seq (mRNA)
miRTools2.0	[107]	centre.bioinformatics.zj.cn/mr2_dev/	1.0, 2013	SA, WS	GUI	miRDeep2, miReap	Quantified	6 tools	Yes	Custom	Yes	No
sRNAtoolbox	[108]	bioinfo5.ugr.es/srnatoolbox/index	1.0, 02/05/2015	WS	GUI	sRNAbench	Quantified	TargetSpy, miRanda, PITA	Yes	edgeR, DESeq, NOISeq	Yes	No

Note: For each toolkit, the table provides references and main features.

the programs, or manually by the user with complementary scripts usually bundled with the prediction software or dedicated tools such as FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) and Trimmomatic [96].

Next, in most cases, clean reads are collapsed into unique sequences and converted to FASTA format to reduce the number of sequences to be aligned in later stages. Further compression of the files is recommended to reduce upload time if a Web-based service is used for the analyses.

MiRNA prediction and expression quantification

After read trimming and filtering, most prediction tools map the reads against known ncRNAs and exclude the aligned reads from following stages. Then, reads are mapped against the genome with a short-read aligner (Table 1), usually Bowtie [97], set to tolerate one or at most two mismatches to consider single-nucleotide polymorphisms or postprocessing base additions. Moreover, short reads can equally align to many genomic loci (multiple mapping reads) because of repetitive regions and paralog genes. In particular, miRNA families and multicopy miRNA precursors are characterized by highly similar genomic sequences. To exclude highly repetitive regions, the reads mapping in more than a fixed number of loci can be discarded and the remaining multiple mapping reads can reasonably be considered representative of miRNAs and are eligible for further processing. However, multiple mapping reads can bias expression quantification of miRNAs derived from multicopy miRNA precursors: counting separately each alignment can overestimate the expression; on the other hand, considering only uniquely mapping reads can lead to the opposite. Some tools, such as miRDeep2, correct expression levels for multiple mappings: a read mapping with equally good alignments against *n* different loci is counted as 1/*n* reads toward the expression of each sequence it aligns against. At this stage, most programs additionally identify and quantify the expression levels of known miRNAs.

Novel miRNA prediction is performed with one of the strategies described before (see section 'MiRNA prediction software tools'). The resulting sequences, genomic coordinates and expression levels of the predicted precursors and matures are output in textual form, usually as table or Browser Extensible Data (BED) (https://genome.ucsc.edu/FAQ/FAQformat.html#format1) files, and often displayed with the support of figures and graphs in PDF or HTML documents. While PDF or HTML results provide the user a report of the predicted miRNAs easy to read and to interpret, results in textual format better suit further processing according to the study objectives.

Downstream analyses: isomiRs, differential expression, target prediction and functional enrichments

Known and novel miRNA isomiRs can be detected if the genomic mapping allows mismatches. Currently, various tools allow isomiR identification and characterization (see the 'IsomiR' column in Tables 1 and 2). There is also software specific for isomiR analysis such as IsomiRID [109] and IsomiR-SEA [110]. In many cases, isomiR characterization could be important to disclose the exact sequences representing the mature miRNAs in the analyzed samples, providing information useful to efficiently design validations and to correctly predict miRNA targets.

Typically, after miRNA detection and quantification, the next step aims at the identification of miRNAs differentially expressed in sample groups (Tables 1 and 2). Several R packages fulfill this purpose, for example edgeR [111, 112], DESeq [113], DESeq2 [114] and/or NOISeq [115].

MiRNA target prediction [116, 117] is a key step to interpret the impact of newly discovered sRNAs on gene expression and modulation of biological processes. Custom target prediction according to the novel miRNA sequence, possibly considering also isomiRs, can be obtained with different methods such as TargetScan [118], miRanda [119], PITA [111] and RNAhybrid [120]. Some of the presented toolkits provide target prediction with one or more of these methods (Table 2). Functional annotation or enrichment analyses of the predicted miRNA targets can be performed with tools such as DAVID [121, 122], EnrichR [122] or g: Profiler [123], considering all the new miRNAs or focusing only the differentially expressed ones. More advanced analyses could link through target prediction both new and known miRNAs to pathways topology to infer the impact of miRNAs on pathway activation [124, 125].

All-in-one analysis toolkits

Most sRNA-seq-based miRNA prediction software estimate expression levels, and some of them also compute differential expression tests. However, there are all-in-one methods designed to perform the different steps of sRNA-seq data analysis (Table 2): from preprocessing of raw reads to differential expression and functional analyses (Figure 2). These general-purpose toolkits integrate the miRNA prediction programs described above, f.i. many of them use the miRDeep2's miRNA prediction module.

All-in-one solutions free the user from possible issues caused by converting files into the formats required by each specific program, but are less flexible and generally offer a smaller range of options and parameter configurations than specific single-purpose software. An exception is the Sepia toolkit [106], whose modular structure allows a high degree of flexibility.

Some toolkits are available as virtual machines (miARMA-Seq [103], CAP-miRSeq [102]) or Docker (https://www.docker.com/) images (miARMA-Seq [103], Sepia [106]), which allow the user to avoid complex installation procedures by providing the software bundled in an environment with all the necessary dependencies already installed, at the cost of more computational resources required than stand-alone applications.

Validation of miRNA predictions

Northern blots and quantitative reverse transcription polymerase chain reaction (qRT-PCR) are the two main approaches for the experimental validation of miRNAs. These methods have a low throughput, which is not adequate to validate hundreds of miRNAs and isomiRs that can be predicted by RNA-seq studies. However, other types of RNA-seq-based experiments can support the annotation of many miRNAs (Table 1). For instance, sRNA sequencing of a cell line or organism before and after the knockdown or knockout of genes of the miRNA biogenesis pathway may assess the biosynthesis of the predicted miRNAs, as the biogenesis of real miRNAs is expected to be affected. In fact, Dicer silencing by RNA interference was used for the validation of miRDeep2 [72] and miRDeep* [84] predictions, while Dicer knockout was used as a control for the validation of MirPlex results [77].

Moreover, cross-linking immunoprecipitation and sequencing (CLIP-seq), which is commonly used to investigate miRNA-transcript interactions [126], can be exploited not only for the validation of putative miRNAs but also to verify their activity. Overlaps between predicted miRNAs and microprocessor complex subunits or Argonaute proteins CLIP-seq tags provide evidence of interactions between the candidate sequence and the microprocessor or RISC complexes. Londin and colleagues [70] used several AGO CLIP-seq data sets, partly publicly available and partly produced in the study, to support the annotation of >2000 newly identified human miRNAs. Further, Friedländer and coworkers [22] used DGCR8, Ago1 and Ago2 CLIP-seq data by to find support for hundreds of putative new miRNAs.

Cross-linking ligation and sequencing of hybrids (CLASH) allows to sequence interacting miRNA-mRNA pairs bound to AGO proteins [127, 128], thus providing direct evidence of the miRNA guiding the RISC onto a target RNA. In the above cited study, Friedländer and coworkers [22] confirmed by CLASH the interactions between a small set of predicted miRNAs and their target mRNAs. However, to our knowledge, no toolkit or platform allowing the evaluation of novel miRNA annotations with CLIP-seq or CLASH data is currently available.

Many studies presenting new miRNA discovery software [72, 82, 84] challenged the older methods. Moreover, Williamson *et al.* [129] evaluated four miRNA prediction tools on seven different data sets, and validated 12 novel miRNAs by qRT-PCR. However, most of these comparisons are now obsolete, and to date, a comprehensive performance evaluation of miRNA prediction methods has not been carried out.

Caveats on miRNA analysis tools

Most miRNAs in miRBase are highly expressed, nontissue- or noncell-type specific and conserved across several genera [28]. MiRNAs with such characteristics are more easily detectable by both bioinformatic and experimental approaches, so they had more chance to be annotated. For instance, methods relying exclusively on conservation information cannot identify species-specific miRNAs. Additionally, tissue- or cell-type-specific miRNAs can be discovered either by genome-based methods, which are affected by low specificity, or using sRNA-seq data of the tissues or cells in which they are expressed. Pooling reads from different sRNA-seq experiments including multiple tissues could be useful to improve detection rate. Recently, Friedländer and colleagues [22] collected a large human sRNA-seq data set (pooling 94 public samples from different cell lines and tissues), and used resampling rarefaction simulations to evaluate miRNA discovery power in relation to sequencing depth and data set number. They demonstrated the advantages of data set pooling (Figure 1C) and also showed that miRNA detection saturation was not reached, indicating that probably many miRNAs remain yet undiscovered even in the human genome. By this approach miRNAs expressed at low levels, down to less than one copy per cell, can be detected, as the combined data better support their read signatures (Figure 1C).

Conversely to pooling, requiring candidate miRNAs to be supported by reads from different biological replicates allows to limit false positives because of technical biases, like single over-amplified RNAs (derived from the PCR step in sRNA-seq experiments) because finding the same sequence in independent data sets provides more robust evidence.

However, some caution should be exercised when selecting data sets for discovery purpose, as including reads from cancer-derived or otherwise abnormal samples can cause the erroneous classification as miRNAs of sequences that are not present or expressed in physiological conditions.

Although miRNA prediction with sRNA-seq data has high sensitivity, it is still biased against some categories of miRNAs including precursors generated by the processing of other ncRNAs. This is because most miRNA prediction software, either rule-based or relying on machine learning techniques, use features typical of canonical miRNA hairpins to evaluate putative precursor candidates. To avoid false positives caused by the presence of reads derived from ncRNA degradation fragments, especially tRNAs, most prediction software filter and exclude all the reads or predicted precursor sequences, which overlap known RNAs, even though some of such loci have been previously linked with the production of biologically active miRNAs [35]. For similar reasons, most miRNA prediction tools discard reads mapping to repetitive sequences (either by the direct masking of repeats or by excluding reads aligned to too many genomic locations). Nonetheless, several miRNAs are known to originate from repetitive regions [6, 21]. Some tools, such as miRDeep2, adopt a conservative choice by flagging the predicted precursors that overlap known ncRNAs and warn the user of the potential false positives. Apart from degradation fragments, reads mapping to known ncRNAs could represent functional products of noncanonical miRNA biogenesis, or other ncRNAs derived from the regulated processing of various ncRNA species. For instance, tRFs (transfer RNA-related fragments) [40, 130], a heterogeneous class of functional sncRNAs generated from the regulated cleavage of tRNAs and tRNA precursors, can be identified with several programs [131]. Further, natural antisense short interfering RNA (nat-siRNA) generated from processing of complementary transcripts in specific conditions can be detected with NATpipe [132] from sRNA-seq reads not assigned to miRNA genes, even in the absence of genome data, grounding on *de novo* assembled transcriptomes.

miRNA genes can also yield microRNA-offset RNA (moRNA) [54, 133–135], which can be differentially expressed in stem [136] and cancer cells [54, 133]. Methods applying narrow models and strict filtering criteria could erroneously discard reads derived from moRNAs [137]. On the contrary, the cleanness of the cut-site between the miRNA and moRNA reinforces the reliability of precursor predictions [138]. The a priori exclusion of reads corresponding to these sRNAs can hamper the identification of functional novel miRNAs and miRNA-like molecules. For this reason, tools not hiding ambiguous elements should be preferred, as they allow manual curation [138].

Conclusions

sRNA-seq by NGS technologies brought a dramatic increase of newly annotated miRNAs in the past 10 years. The development of bioinformatics methods to characterize novel sRNAs is still an active field of research, which now can benefit from a wealth of sRNA-seq data, as it was demonstrated by the reanalysis of publicly available data sets that made possible the post hoc assessment of previous miRBase miRNA annotations and the definition of a high confidence miRNA set [20].

Given the large number of tools available, researchers willing to analyze data from sRNA-seq experiments may wonder which method is the most appropriate for their study. For instance, read signature evaluation approaches that are based on machine learning algorithms are more efficient when applied to animals with a large number of miRNA annotations and/or with well-annotated closely related species. Such knowledge base can enhance the training of the model and improve quality of predictions. Yet, this approach is biased toward the discovery of miRNAs similar to those already known because negative sets for training cannot be easily established [61, 139]. Moreover, there is no consensus on which machine learning approach works best [139, 140]. In contrast, rule-based read signature evaluation approaches can be applied to any species with a reference genome. Among these methods, the most cited in literature is miRDeep2, which was incorporated into several all-in-one toolkits.

If the analysis of ncRNAs other than miRNAs is relevant for the study, read profile clustering methods like FlaiMapper [88], DARIO [91] or miRDBA [92] can identify known and novel ncRNAs of several classes including tRNAs, scRNAs, snoRNAs and snRNAs. Additionally, CoRAL [80, 94], which relies on a machine learning algorithm for the evaluation of read signatures, allows the prediction of five other classes of ncRNAs (lincRNAs, scRNAs, C/D box snoRNAs, snRNAs and transposon-derived snRNAs) as well as miRNAs. The miR&moRe pipeline [133] can predict novel miRNA* from known precursors, as well as isomiRs and moRNAs, and can be combined with miRDeep2 to identify miRNA, moRNAs and isomiRs from novel precursors [134].

Users with limited computational resources can use Web-based services, or software available as Galaxy [141] tools or workflows, such as BlockClust [86]. These tools are also the most user-friendly, as they are configured through a graphical user interface and do not need to be installed on the user's machine. Furthermore, they allow the recording of the parameters, inputs and versions of every tool, making analyses easily sharable and reproducible. However, Web-based tools are often limited to few predefined species and model organisms. For instance, the DARIO Web server allows the user to analyze data from six model organisms (human, rhesus macaque, mouse, fruit fly, nematode and zebrafish). Unfortunately, Web-based services limit the amount of data that can be uploaded, preventing the analysis of sRNA-seq data obtained by data set pooling or produced at high sequencing depths, which are the most informative for miRNA discovery and are becoming the standard in current research projects.

Key Points

- MiRNA annotation is clearly incomplete in animals, especially in nonmodel species, but also in humans and well-studied model organisms.
- sRNA-seq data analysis allows the discovery and the characterization of hundreds novel miRNAs.
- Several tools for miRNA prediction in animals are available, each one presenting bias or limitations that should be taken into account.
- Typical miRNA prediction and analysis workflow considers several steps, from preparatory data processing to diverse analyses downstream miRNA detection and quantification.
- Pooling different sRNA-seq data sets can improve detection power in discovery experiments, and other high-throughput sequencing data (such as CLIP-seq or CLASH) can validate novel miRNAs.

Funding

We thank for financial support to SB Fondazione Cassa di Risparmio di Padova e Rovigo, AIRC (Italian Association for Cancer Research) and University of Padova.

References

1. Leung AK. The whereabouts of microRNA actions: cytoplasm and beyond. *Trends Cell Biol* 2015;**25**(10):601–10.
2. Catalanotto C, Cogoni C, Zardo G. MicroRNA in control of gene expression: an overview of nuclear functions. *Int J Mol Sci* 2016;**17**:1712–29.
3. Kovanda A, Režen T, Rogelj B. MicroRNA in skeletal muscle development, growth, atrophy, and disease. *Wiley Interdiscip Rev RNA* 2014;**5**:509–25.
4. Abernathy DG, Yoo AS. MicroRNA-dependent genetic networks during neural development. *Cell Tissue Res* 2015;**359**(1):179–85.
5. Johanson TM, Skinner JP, Kumar A, *et al*. The role of microRNAs in lymphopoiesis. *Int J Hematol* 2014;**100**(3): 246–53.
6. Berezikov E. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* 2011;**12**(12):846–60.
7. Bhayani MK, Calin GA, Lai SY. Functional relevance of miRNA sequences in human disease. *Mutat Res* 2012;**731**: 14–19.
8. Chan B, Manley J, Lee J, *et al*. The emerging roles of microRNAs in cancer metabolism. *Cancer Lett* 2015;**356**(2): 301–8.
9. Hata A, Lieberman J. Dysregulation of microRNA biogenesis and gene silencing in cancer. *Sci Signal* 2015;**8**(368):re3.
10. Rupaimoole R, Calin GA, Lopez-Berestein G, *et al*. miRNA deregulation in cancer cells and the tumor microenvironment. *Cancer Discov* 2016;**6**(3):235–46.
11. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. *Trends Mol Med* 2014;**20**: 460–9.
12. Schmidt MF. Drug target miRNAs: chances and challenges. *Trends Biotechnol* 2014;**32**(11):578–85.
13. Xia T, Li J, Cheng H, *et al*. Small-molecule regulators of MicroRNAs in biomedicine. *Drug Dev Res* 2015;**76**:375–81.
14. Banwait JK, Bastola DR. Contribution of bioinformatics prediction in microRNA-based cancer therapeutics. *Adv Drug Deliv Rev* 2015;**81**:94–103.
15. Perilli L, Vicentini C, Agostini M, *et al*. Circulating miR-182 is a biomarker of colorectal adenocarcinoma progression. *Oncotarget* 2014;**5**(16):6611–19.
16. Fatima F, Nawaz M. Vesiculated long non-coding RNAs: offshore packages deciphering trans-regulation between cells, cancer progression and resistance to therapies. *Noncoding RNA* 2017;**3**:10–33.
17. Wu N, Zhu Q, Chen B, *et al*. High-throughput sequencing of pituitary and hypothalamic microRNA transcriptome associated with high rate of egg production. *BMC Genomics* 2017; **18**:255–68.
18. Gajigan AP, Conaco C. A microRNA regulates the response of corals to thermal stress. *Mol Ecol* 2017;**26**:3472–83.
19. Ioannidis J, Donadeu FX. Circulating miRNA signatures of early pregnancy in cattle. *BMC Genomics* 2016;**17**:184–96.
20. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;**42**:D68–73.
21. Axtell MJ, Westholm JO, Lai EC. Vive la différence: biogenesis and evolution of microRNAs in plants and animals. *Genome Biol* 2011;**12**:221–34.
22. Friedländer MR, Lizano E, Houben AJS, *et al*. Evidence for the biogenesis of more than 1, 000 novel human microRNAs. *Genome Biol* 2014;**15**(4):R57.
23. McCormick KP, Willmann MR, Meyers BC. Experimental design, preprocessing, normalization and differential

expression analysis of small RNA sequencing experiments. *Silence* 2011;**2**:2–21.

24. Raabe CA, Tang TH, Brosius J, *et al*. Biases in small RNA deep sequencing data. *Nucleic Acids Res* 2014;**42**(3):1414–26.
25. Witwer KW, Halushka MK. Toward the promise of microRNAs—enhancing reproducibility and rigor in microRNA research. *RNA Biol* 2016;**13**(11):1103–16.
26. Zhao S, Gordon W, Du S, *et al*. QuickMIRSeq: a pipeline for quick and accurate quantification of both known miRNAs and isomiRs by jointly processing multiple samples from microRNA sequencing. *BMC Bioinformatics* 2017;**18**:180–914.
27. Huang PJ, Liu YC, Lee CC, *et al*. DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res* 2010;**38**:W385–91.
28. Kang W, Friedländer MR. Computational prediction of miRNA genes from small RNA sequencing data. *Front Bioeng Biotechnol* 2015;**3**:7.
29. Rajendiran A, Chatterjee A, Pan A. Computational approaches and related tools to identify MicroRNAs in a species: a bird's eye view. *Interdiscip Sci* 2017:1–20. doi: 10.1007/s12539-017-0223-x.
30. Gomes CP, Cho JH, Hood L, *et al*. A review of computational tools in microRNA discovery. *Front Genet* 2013;**4**:81.
31. Gupta R, Davuluri RV. Bioinformatics approaches to the study of MicroRNAs. In: M Fabbri (eds) *Non-Coding RNAs and Cancer*. New York, NY: Springer, 2013, 165–245.
32. Akhtar MM, Micolucci L, Islam MS, *et al*. Bioinformatic tools for microRNA dissection. *Nucleic Acids Res* 2016;**44**:24–44.
33. Ha M, Narry Kim V. Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol* 2014;**15**(8):509–24.
34. Abdelfattah AM, Park C, Choi MY. Update on non-canonical microRNAs. *Biomol Concepts* 2014;**5**(4):275–87.
35. Maute RL, Dalla-Favera R, Basso K. RNAs with multiple personalities. *Wiley Interdiscip Rev RNA* 2014;**5**:1–13.
36. Curtis HJ, Sibley CR, Wood MJA. Mirtrons, an emerging class of atypical miRNA. *Wiley Interdiscip Rev RNA* 2012;**3**:617–32.
37. Scott MS, Ono M. From snoRNA to miRNA: dual function regulatory non-coding RNAs. *Biochimie* 2011;**93**(11):1987–92.
38. Falaleeva M, Stamm S. Processing of snoRNAs as a new source of regulatory non-coding RNAs. *Bioessays* 2013;**35**(1): 46–54.
39. Anderson P, Ivanov P. tRNA fragments in human health and disease. *FEBS Lett* 2014;**588**(23):4297–304.
40. Venkatesh T, Suresh PS, Tsutsumi R. tRFs: miRNAs in disguise. *Gene* 2016;**579**(2):133–8.
41. Seok H, Ham J, Jang ES, *et al*. MicroRNA target recognition: insights from transcriptome-wide non-canonical interactions. *Mol Cells* 2016;**39**:375–81.
42. Izaurralde E. Breakers and blockers—miRNAs at work. *Science* 2015;**349**(6246):380–2.
43. Jonas S, Izaurralde E. Towards a molecular understanding of microRNA-mediated gene silencing. *Nat Rev Genet* 2015;**16**: 421–33.
44. Iwakawa HO, Tomari Y. The functions of MicroRNAs: mRNA decay and translational repression. *Trends Cell Biol* 2015; **25**(11):651–65.
45. Sonda N, Simonato F, Peranzoni E, *et al*. miR-142-3p prevents macrophage differentiation during cancer-induced myelopoiesis. *Immunity* 2013; **38**(6):1236–49.
46. Quann K, Jing Y, Rigoutsos I. Post-transcriptional regulation of BRCA1 through its coding sequence by the miR-15/107 group of miRNAs. *Front Genet* 2015;**6**:242.
47. Neilsen CT, Goodall GJ, Bracken CP. IsomiRs—the overlooked repertoire in the dynamic microRNAome. *Trends Genet* 2012;**28**(11):544–9.
48. Guo L, Chen F. A challenge for miRNA: multiple isomiRs in miRNAomics. *Gene* 2014;**544**(1):1–7.
49. Fernandez-Valverde SL, Taft RJ, Mattick JS. Dynamic isomiR regulation in Drosophila development. *Rna* 2010;**16**(10): 1881–8.
50. Morin RD, O'Connor MD, Griffith M, *et al*. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res* 2008;**18**(4): 610–21.
51. Penso-Dolfin L, Swofford R, Johnson J, *et al*. An improved microRNA annotation of the canine genome. *PLoS One* 2016; **11**(4):e0153453.
52. Azuma-Mukai A, Oguri H, Mituyama T, *et al*. Characterization of endogenous human argonautes and their miRNA partners in RNA silencing. *Proc Natl Acad Sci USA* 2008;**105**:7964–9.
53. Tan GC, Chan E, Molnar A, *et al*. 5′ isomiR variation is of functional and evolutionary importance. *Nucleic Acids Res* 2014; **42**(14):9424–35.
54. Guglielmelli P, Bisognin A, Saccoman C, *et al*. Small RNA sequencing uncovers new miRNAs and moRNAs differentially expressed in normal and primary myelofibrosis CD34+ Cells. *PLoS One* 2015;**10**(10):e0140445.
55. Lagos-Quintana M, Rauhut R, Meyer J, *et al*. New microRNAs from mouse and human. *RNA* 2003;**9**(2):175–9.
56. Lau NC, Lim LP, Weinstein EG, *et al*. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* 2001;**294**(5543):858–62.
57. Bentwich I. Prediction and validation of microRNAs and their targets. *FEBS Lett* 2005;**579**(26):5904–10.
58. Lai EC, Tomancak P, Williams RW, *et al*. Computational identification of Drosophila microRNA genes. *Genome Biol* 2003; **4**(7):R42.
59. Lim LP, Lau NC, Weinstein EG, *et al*. The microRNAs of *Caenorhabditis elegans*. *Genes Dev* 2003;**17**(8):991–1008.
60. Kadri S, Hinman V, Benos PV. HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. *BMC Bioinformatics* 2009;**10(Suppl 1)**:S35.
61. Wei L, Liao M, Gao Y, *et al*. Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set. *IEEE/ACM Trans Comput Biol Bioinform* 2013;**11**: 192–201.
62. Wang X, Zhang J, Li F, *et al*. MicroRNA identification based on sequence and structure alignment. *Bioinformatics* 2005; **21**(18):3610–14.
63. Zhang Y, Yang Y, Zhang H, *et al*. Prediction of novel premicroRNAs with high accuracy through boosting and SVM. *Bioinformatics* 2011;**27**(10):1436–7.
64. Jha A, Chauhan R, Mehra M, *et al*. miR-BAG: bagging based identification of microRNA precursors. *PLoS One* 2012;**7**(9): e45782.
65. Kuenne C, Preussner J, Herzog M, *et al*. MIRPIPE: quantification of microRNAs in niche model organisms. *Bioinformatics* 2014;**30**(23):3412–13.
66. Higashi S, Fournier C, Gautier C, *et al*. Mirinho: an efficient and general plant and animal pre-miRNA predictor for genomic and deep sequencing data. *BMC Bioinformatics* 2015; **16**:179.
67. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 2011;**39**:D152–7.
68. Pacholewska A, Mach N, Mata X, *et al*. Novel equine tissue miRNAs and breed-related miRNA expressed in serum. *BMC Genomics* 2016;**17**:831–46.

69. Wake C, Labadorf A, Dumitriu A, *et al*. Novel microRNA discovery using small RNA sequencing in post-mortem human brain. *BMC Genomics* 2016;**17**:776–85.
70. Londin E, Loher P, Telonis AG, *et al*. Analysis of 13 cell types reveals evidence for the expression of numerous novel primate- and tissue-specific microRNAs. *Proc Natl Acad Sci USA* 2015;**112**:E1106–15.
71. Friedländer MR, Chen W, Adamidi C, *et al*. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* 2008;**26**(4):407–15.
72. Friedländer MR, Mackowiak SD, Li N, *et al*. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 2012;**40**(1): 37–52.
73. Tyler DM, Okamura K, Chung WJ, *et al*. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes Dev* 2008;**22**(1):26–36.
74. Stark A, Bushati N, Jan CH, *et al*. A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands. *Genes Dev* 2008;**22**(1):8–13.
75. Okamura K, Lai EC. Endogenous small interfering RNAs in animals. *Nat Rev Mol Cell Biol* 2008;**9**(9):673–8.
76. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, *et al*. ViennaRNA package 2.0. *Algorithms Mol Biol* 2011;**6**:26–40.
77. Mapleson D, Moxon S, Dalmay T, *et al*. MirPlex: a tool for identifying miRNAs in high-throughput sRNA datasets without a genome. *J Exp Zool B Mol Dev Evol* 2013;**320**(1):47–56.
78. Jha A, Shankar R, Zhao Z. miReader: discovering novel miRNAs in species without sequenced genome. *PLoS One* 2013;**8**(6):e66857.
79. Hackenberg M, Rodriguez-Ezpeleta N, Aransay AM. miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res* 2011;**39**:W132–8.
80. Leung YY, Ryvkin P, Ungar LH, *et al*. CoRAL: predicting noncoding RNAs from small RNA-sequencing data. *Nucleic Acids Res* 2013;**41**(14):e137.
81. Qian K, Auvinen E, Greco D, *et al*. miRSeqNovel: an R based workflow for analyzing miRNA sequencing data. *Mol Cell Probes* 2012;**26**:208–11.
82. Hansen TB, Venø MT, Kjems J, *et al*. miRdentify: high stringency miRNA predictor identifies several novel animal miRNAs. *Nucleic Acids Res* 2014;**42**(16):e124.
83. Mathelier A, Carbone A. MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics* 2010;**26**(18):2226–34.
84. An J, Lai J, Lehman ML, *et al*. miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data. *Nucleic Acids Res* 2013;**41**(2):727–37.
85. Barturen G, Rueda A, Hamberg M, *et al*. sRNAbench: profiling of small RNAs and its sequence variants in single or multi-species high-throughput experiments. *Methods Next Gener Seq* 2014;**1**:21–31.
86. Videm P, Rose D, Costa F, *et al*. BlockClust: efficient clustering and classification of non-coding RNAs from short read RNA-seq profiles. *Bioinformatics* 2014;**30**(12):i274–82.
87. Langenberger D, Pundhir S, Ekstrøm CT, *et al*. deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns. *Bioinformatics* 2012;**28**(1):17–24.
88. Hoogstrate Y, Jenster G, Martens-Uzunova ES. FlaiMapper: computational annotation of small ncRNA-derived fragments using RNA-seq high-throughput data. *Bioinformatics* 2015;**31**(5):665–73.
89. Olvedy M, Scaravilli M, Hoogstrate Y, *et al*. A comprehensive repertoire of tRNA-derived fragments in prostate cancer. *Oncotarget* 2016;**7**(17):24766–77.
90. Martens-Uzunova ES, Hoogstrate Y, Kalsbeek A, *et al*. C/D-box snoRNA-derived RNA production is associated with malignant transformation and metastatic progression in prostate cancer. *Oncotarget* 2015;**6**(19):17430–44.
91. Fasold M, Langenberger D, Binder H, *et al*. DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res* 2011;**39**:W112–17.
92. Pundhir S, Gorodkin J. MicroRNA discovery by similarity search to a database of RNA-seq profiles. *Front Genet* 2013;**4**: 133–46.
93. O N Lopes Id, Schliep A, de L F de Carvalho AP. Automatic learning of pre-miRNAs from different species. *BMC Bioinformatics* 2016;**17**:224–42.
94. Ryvkin P, Leung YY, Ungar LH, *et al*. Using machine learning and high-throughput RNA sequencing to classify the precursors of small non-coding RNAs. *Methods* 2014;**67**(1):28–35.
95. Langenberger D, Bermudez-Santana CI, Stadler PF, *et al*. Identification and classification of small RNAs in transcriptome sequence data. *Pac Symp Biocomputing* 2010;**15**:80–7.
96. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**(15): 2114–20.
97. Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* 2010;**Chapter 11**:Unit 11.7.
98. Beckers M, Mohorianu I, Stocks M, *et al*. Comprehensive processing of high-throughput small RNA sequencing data including quality checking, normalization, and differential expression analysis using the UEA sRNA Workbench. *RNA* 2017;**23**(6):823–35.
99. Zhao W, Liu W, Tian D, *et al*. wapRNA: a web-based application for the processing of RNA sequences. *Bioinformatics* 2011;**27**(21):3076–7.
100. Müller S, Rycak L, Winter P, *et al*. omiRas: a web server for differential expression analysis of miRNAs derived from small RNA-Seq data. *Bioinformatics* 2013;**29**(20):2651–2.
101. Yuan T, Huang X, Dittmar RL, *et al*. eRNA: a graphic user interface-based tool optimized for large data analysis from high-throughput RNA sequencing. *BMC Genomics* 2014;**15**: 176.
102. Sun Z, Evans J, Bhagwate A, *et al*. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. *BMC Genomics* 2014;**15**:423–33.
103. Andrés-León E, Núñez-Torres R, Rojas AM. miARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis. *Sci Rep* 2016;**6**:25749.
104. Capece V, Garcia Vizcaino JC, Vidal R, *et al*. Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics* 2015;**31**(13):2205–7.
105. Giurato G, De Filippo MR, Rinaldi A, *et al*. iMir: an integrated pipeline for high-throughput analysis of small non-coding RNA data obtained by smallRNA-Seq. *BMC Bioinformatics* 2013;**14**:362.
106. Icay K, Chen P, Cervera A, *et al*. SePIA: RNA and small RNA sequence processing, integration, and analysis. *BioData Min* 2016;**9**:20–38.
107. Wu J, Liu Q, Wang X, *et al*. mirTools 2.0 for non-coding RNA discovery, profiling, and functional annotation based on high-throughput sequencing. *RNA Biol* 2013;**10**(7):1087–92.
108. Rueda A, Barturen G, Lebrón R, *et al*. sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res* 2015;**43**(W1):W467–73.

109. de Oliveira LF, Christoff AP, Margis R. isomiRID: a framework to identify microRNA isoforms. *Bioinformatics* 2013;**29**(20): 2521–3.
110. Urgese G, Paciello G, Acquaviva A. isomiR-SEA: an RNA-Seq analysis tool for miRNAs/isomiRs expression level profiling and miRNA-mRNA interaction sites evaluation. *BMC Bioinformatics* 2016;**17**:148–61.
111. Kertesz M, Iovino N, Unnerstall U, *et al*. The role of site accessibility in microRNA target recognition. *Nat Genet* 2007; **39**(10):1278–84.
112. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**(1):139–40.
113. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106–18.
114. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 2014;**15**:550–71.
115. Tarazona S, Furió-Tarí P, Turrà D, *et al*. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Res* 2015;**43**:e140.
116. Fan X, Kurgan L. Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Brief Bioinform* 2015;**16**:780–94.
117. Riffo-Campos ÁL, Riquelme I, Brebi-Mieville P. Tools for sequence-based miRNA target prediction: what to choose? *Int J Mol Sci* 2016;**17**:1987–2005.
118. Lewis BP, Shih IH, Jones-Rhoades MW, *et al*. Prediction of mammalian MicroRNA targets. *Cell* 2003;**115**(7):787–98.
119. John B, Enright AJ, Aravin A, *et al*. Human MicroRNA targets. *PLoS Biol* 2004;**2**(11):e363.
120. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. *RNA* 2004;**10**(10):1507–17.
121. Huang DW, Sherman BT, Tan Q. DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* 2007;**35**:W169–75.
122. Kuleshov MV, Jones MR, Rouillard AD, *et al*. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016; **44**(W1):W90–7.
123. Reimand J, Arak T, Vilo J. g: profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res* 2011;**39**:W307–15.
124. Calura E, Pizzini S, Bisognin A, *et al*. A data-driven network model of primary myelofibrosis: transcriptional and post-transcriptional alterations in CD34+ cells. *Blood Cancer J* 2016;**6**(6):e439.
125. Calura E, Bisognin A, Manzoni M, *et al*. Disentangling the microRNA regulatory milieu in multiple myeloma: integrative genomics analysis outlines mixed miRNA-TF circuits and pathway-derived networks modulated in t(4; 14) patients. *Oncotarget* 2016;**7**(3):2367–78.
126. Mittal N, Zavolan M. Seq and CLIP through the miRNA world. *Genome Biol* 2014;**15**(1):202.
127. Helwak A, Tollervey D. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat Protoc* 2014;**9**(3):711–28.
128. Broughton JP, Pasquinelli AE. A tale of two sequences: microRNA-target chimeric reads. *Genet Sel Evol* 2016;**48**:31–8.
129. Williamson V, Kim A, Xie B, *et al*. Detecting miRNAs in deep-sequencing data: a software performance comparison and evaluation. *Brief Bioinform* 2013;**14**(1):36–45.
130. Kumar P, Kuscu C, Dutta A. Biogenesis and function of transfer RNA-related fragments (tRFs). *Trends Biochem Sci* 2016; **41**(8):679–89.
131. Xu WL, Yang Y, Wang YD, *et al*. Computational approaches to tRNA-derived small RNAs. *Noncoding RNA* 2017;**3**:2–14.
132. Yu D, Meng Y, Zuo Z, *et al*. NATpipe: an integrative pipeline for systematical discovery of natural antisense transcripts (NATs) and phase-distributed nat-siRNAs from de novo assembled transcriptomes. *Sci Rep* 2016;**6**:21666.
133. Bortoluzzi S, Bisognin A, Biasiolo M, *et al*. Characterization and discovery of novel miRNAs and moRNAs in JAK2V617F-mutated SET2 cells. *Blood* 2012;**119**(13):e120–30.
134. Gaffo E, Zambonelli P, Bisognin A, *et al*. miRNome of Italian large white pig subcutaneous fat tissue: new miRNAs, isomiRs and moRNAs. *Anim Genet* 2014;**45**:685–98.
135. Shi W, Hendrix D, Levine M, *et al*. A distinct class of small RNAs arises from pre-miRNA–proximal regions in a simple chordate. *Nat Struct Mol Biol* 2009;**16**(2):183–9.
136. Asikainen S, Heikkinen L, Juhila J, *et al*. Selective microRNA-Offset RNA expression in human embryonic stem cells. *PLoS One* 2015;**10**(3):e0116668.
137. Bortoluzzi S, Biasiolo M, Bisognin A. MicroRNA–offset RNAs (moRNAs): by-product spectators or functional players? *Trends Mol Med* 2011;**17**(9):473–4.
138. Hendrix D, Levine M, Shi W. miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol* 2010;**11**(4): R39.
139. Saçar Demirci MD, Allmer J. Delineating the impact of machine learning elements in pre-microRNA detection. *PeerJ* 2017;**5**:e3131.
140. Jiang L, Zhang J, Xuan P, *et al*. BP neural network could help improve pre-miRNA identification in various species. *Biomed Res Int* 2016;**2016**:9565689.
141. Afgan E, Baker D, van den Beek M, *et al*. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 2016;**44**(W1): W3–W10.