2018

miRDis: a Web tool for endogenous and exogenous microRNA discovery based on deep-sequencing data analysis

Hanyuan Zhang
University of Nebraska-Lincoln

Bruno Vieira Resende e Silva
University of Nebraska-Lincoln, bsilva2@unl.edu

Juan Cui
University of Nebraska-Lincoln, jcui@unl.edu

OXFORD

Briefings in Bioinformatics, 19(3), 2018, 415–424

doi: 10.1093/bib/bbw140
Advance Access Publication Date: 10 January 2017
Paper

miRDis: a Web tool for endogenous and exogenous microRNA discovery based on deep-sequencing data analysis

Hanyuan Zhang, Bruno Vieira Resende e Silva and Juan Cui

Corresponding author: Juan Cui, Department of Computer Science and Engineering, Systems Biology and Biomedical Informatics (SBBI) Laboratory, University of Nebraska-Lincoln, Lincoln, NE 68588, USA. Tel.: 402-472-5023; Fax: 402-472-7767; E-mail: jcui@unl.edu

Abstract

Small RNA sequencing is the most widely used tool for microRNA (miRNA) discovery, and shows great potential for the efficient study of miRNA cross-species transport, i.e., by detecting the presence of exogenous miRNA sequences in the host species. Because of the increased appreciation of dietary miRNAs and their far-reaching implication in human health, research interests are currently growing with regard to exogenous miRNAs bioavailability, mechanisms of cross-species transport and miRNA function in cellular biological processes. In this article, we present microRNA Discovery (miRDis), a new small RNA sequencing data analysis pipeline for both endogenous and exogenous miRNA detection. Specifically, we developed and deployed a Web service that supports the annotation and expression profiling data of known host miRNAs and the detection of novel miRNAs, other noncoding RNAs, and the exogenous miRNAs from dietary species. As a proof-of-concept, we analyzed a set of human plasma sequencing data from a milk-feeding study where 225 human miRNAs were detected in the plasma samples and 44 show elevated expression after milk intake. By examining the bovine-specific sequences, data indicate that three bovine miRNAs (bta-miR-378, -181* and -150) are present in human plasma possibly because of the dietary uptake. Further evaluation based on different sets of public data demonstrates that miRDis outperforms other state-of-the-art tools in both detection and quantification of miRNA from either animal or plant sources. The miRDis Web server is available at: http://sbbi.unl.edu/miRDis/index.php.

Key words: MicroRNA sequencing; endogenous microRNA; exogenous microRNA; microRNA expression; novel microRNA

Introduction

MicroRNAs (miRNAs) are a class of functionally important noncoding RNAs that play an important role in posttranscription regulation via destabilizing messenger RNAs (mRNAs) or preventing their translation [1, 2]. They have been long considered synthesized endogenously until recent studies reported that animal can acquire exogenous miRNA through dietary intake [3]. Specially, exogenous miRNA sequences from plant (e.g. rice and honeysuckle) and animal (cow milk and egg) have been detected in the sera and tissues of animals and human [3–6], and the biogenesis and function of such exogenous miRNAs are evidently health related [7–10]. With increasingly soared research enthusiasm on dietary intervention, miRNA functional study has spread out from intracellular posttranscription to

Hanyuan Zhang is a graduate student in the Department of Computer Science and Engineering at University of Nebraska-Lincoln. His research interests include bioinformatics, computational genomics and Web server development.
Bruno Vieira Resende e Silva is a graduate student in the Department of Computer Science and Engineering at University of Nebraska-Lincoln. His research interests include bioinformatics, data mining and Web server development.
Juan Cui is an assistant professor in the Department of Computer Science and Engineering at University of Nebraska-Lincoln. Her research interests include computational biology, biomedical informatics and data mining.
Submitted: 19 September 2016; **Received (in revised form):** 7 December 2016

extracellular signaling, with particular focus on the implication in human health [11], whereas the first key step remains to be the reliable detection of exogenous miRNA transport using next-generation sequencing (NGS). Investigators have detected numerous dietary, nonhuman miRNAs in 6.8 billion sequenced short sequences (reads) from 528 human samples [12] and identified 50 plant-borne miRNAs in human plasma [13] by using small noncoding RNA sequencing. However, the challenges of exogenous miRNA discovery in terms of fast multi-genome read alignment (mapping), cross-species sequence comparison and particularly the differentiation of exogenous sequences that share subtle (or no) difference with their homologs in host species have posed widespread concerns on existing analysis.

During the past decade, numerous tools have been developed for miRNA sequencing data analysis including miRDeep2 [14], CAP-miRSeq [15], DSAP [16], DARIO [17], omiRas [18], sRNAbench [19], ShortStack [20], miRDeep-P [21] and miRPREFeR [22], where the latter two are specifically designed for plant miRNA. These existing tools focus on miRNA expression profiling and novel endogenous miRNA discovery and a few offer downstream analysis on differential expression, miRNA targets and pathway enrichment [16, 18, 23]. Most tools have at least one of such problems as high false discovery rate (FDR), long running time or nonintuitive to use. For example, because of the common concern that multi-genome comparison normally requires significant long computing time, Web-based tools such as DARIO [17], sRNAbench [23] and Chimira [24] limit the user submission by one sample per job and/or file upload size <500 MB or 1.6 GB, while the well-designed stand-alone packages, e.g. CAP-miRSeq [15], require mandatory configuration on local clusters or cloud to process multiple jobs in parallel, which might be challenging for Wet-laboratory users. More importantly, none of them are designed specifically for exogenous miRNA detection. It thus becomes highly desirable to have a user-friendly Web-based tool for both endogenous and exogenous miRNA analysis based on sequencing data.

Empowered by NGS analysis, a common procedure to discover endogenous miRNAs in a certain species and quantify their expression is to map all sequencing reads to the known miRNA sequences archived in the public databases such as miRBase [25], Refseq [26] and Rfam [27] and then annotate them based on sequence similarity. Intuitively, we could do the same for exogenous miRNA analysis, which however will encounter inherent challenges. First, there are certain discrepancies between sequences of miRBase annotation and actual expressed miRNAs because of single-nucleotide polymorphisms (SNPs) and small insertions and deletions [28–31], which make the detection of miRNA isoforms and exogenous miRNAs challenging. Second, the possible novel miRNAs in the host could also be the confounders of exogenous sequences when they are similar. For instance, the existing rule in miRdeep2 [14] that determines a novel miRNA based on the putative expression of both 5′ and 3′ mature miRNAs may lead to conflicts when screening for exogenous candidates.

In this study, we designed an analytical platform to tackle the aforementioned challenges through the following strategies: (i) the whole NGS reads will be first divided into two pools of endogenous and exogenous reads, respectively, according to their mapping status to the genomes of the host and other dietary species with SNPs being considered. (ii) The capacity of the entire sequence analysis pipeline will be significantly advanced by compressing the identical reads from different samples into a sequence tag with complete index of source sample and corresponding read count. In this way, the consensus sequences of the mapped regions are annotated using all reads across multiple samples instead reads from a single sample to ensure a higher recall rate. (iii) Repeat-deprived reads may represent miRNAs coming from repetitive sequences such as transposable elements [32, 33] and will be mapped to multiple genome regions. In this study, we will keep all mapped regions that hold a possible stable stem-loop RNA structure or include unique mapped reads. This pipeline also integrates multiple analytical functions such as sequence alignment (BLAST), precursor structure prediction (RNAfold [34]) and RNA homology search (infernal [35] based on the covariance model (CM) in Rfam [27], where the latest version supports 100-fold faster comparison). To annotate each mapped region, we will use the whole-genome annotation coupled with precursor structure prediction, as opposed to using merely known mature miRNA sequence. Based on all these considerations, we developed a Web service, microRNA discovery (miRDis) based on deep sequencing data to accomplish the proposed pipeline.

Materials and methods

Read preprocess and mapping

First, the sequencing data quality control (QC) tool, FASTQC [36], was integrated into the pipeline to generate QC report. To eliminate the low-quality reads and adaptors, Cutadapt [37] was included by using the user input 3′-adapter sequence or the auto-detected adapter by a wrapper program we developed. Overrepresented sequences (e.g. abundant miRNAs (miR-486-5p) in human blood, 1–48% [38]) are specially treated to avoid bias in adapter detection.

UniqRead sets

The same reads from different samples were collapsed into a uniqRead set and stored in FASTA format with a tag symbol. For example, in '>seq_366_len_22_x21871;1:21857;2:8;3:3;4:3', 'seq_366' is the unique ID of the uniqRead set; 'len_22' indicates the read length; 'x21871' indicates the total count of the reads; and the rest lists the counts in every individual sample. By considering all samples simultaneously, such design is expected to reduce computational load and render higher confidence on the detected reads.

Mapping

The genome annotation on 13 types of animal and plant is downloaded from the Ensemble database [39]. All unique collapsed reads are mapped to the genomes of the host or dietary species by bowtie [40], and the mapped regions are identified using BEDtools [41]. For reads that have more than one mapped region, we assign them to the loci that have more unique reads or more stable secondary structure inferred by RNAfold (Figure 1A). The BED file covers information about all mapped reads and the depth in every single position.

Annotation of the genome mapped regions

After mapping, the consensus sequence from each mapped region is extracted through the following analysis: all peaks [single positions with the highest depth (no less than 5) within a consecutive mapped region] are examined. The region that is extended from each peak toward both sides and ends where the depth is >3-fold lower than the peak becomes an expressed region (Figure 1B). The consensus sequence retrieved from such expressed region will remain for further annotation to be

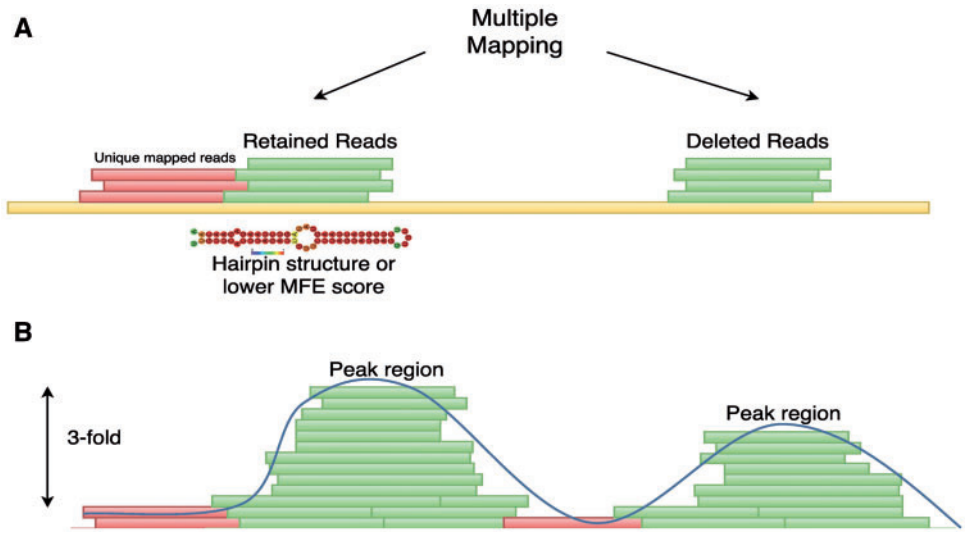

Figure 1. The schematic illustrations of the (**A**) assignment of the multi-mapping reads to a single mapped locus according to the read composition and/or RNA structure prediction, and (**B**) detection of the mapped regions after mapping.

Table 1. Rules applied to annotate the expressed regions as known mature miRNA, novel miRNA and other noncoding RNA, based on the sequence comparison and database annotation

Categories	Type	Sequence similarity			Database annotation	
		I_{mature}	$I_{precursor}$	Rfam (*e*-value)	MirBase	Ensemble
Known	Mature miRNA	=100	–	–	Known mature	–
	IsoMiRs	≥85 and <100	–	–	Known mature	–
	MiR-precursor	<85	>50	–	Known precursor	–
Novel	Novel-miRNA1	–	–	<1E-5	–	Predicted miRNA
	Novel-miRNA2	>40	–	<1E-5	–	Not annotated
Others	Other ncRNA1	–	–	–	–	Noncoding RNA
	Other ncRNA2	–	–	<1E-5	–	Not annotated

categorized into known mature miRNA (either endogenous or exogenous), novel miRNA and other noncoding RNA (rules listed in Table 1).

Known miRNAs

First, the expressed regions that are in accordance with the mature miRNAs annotation on genome coordinate and sequence similar are considered known miRNAs. To evaluate the confidence, we define a new term, mature information (I_{mature}), based on the sequence alignment between the annotated mature sequence and the expressed sequence as follows:

$$I_{mature} = \frac{Identity * 100}{\text{length (sequence of the expressed region)}}.$$

An expressed region with mature information $I = 100$ indicates a perfectly matched mature miRNA, while a lower I (<100 and ≥85) may indicate isomiRs that have one sequence variation or one extended base compared with its mature sequence. The miR-precursor type refers to the sequences that are identical to part of a known precursor but not reported as mature sequence in the databases. In addition, we also used the consensus sequence from each mapped region and their flanking sequence to identify new miRNA precursors through structure prediction using RNAfold [34] and structure similarity comparison using infernal based on Rfam CM [27, 35].

An exogenous miRNA sequence can be determined based on the following scenarios: (1) if the expressed region along with their flanking sequence show a better match with known miRNA sequence in a dietary species compared with the host (e.g. the extended sequence on either side matched to dietary genome but not human, although the expressed region may be identical in both genomes, or the similarity is higher in dietary species versus human), (2) the expressed region is corresponding to a known miRNA of the dietary species but unrelated to any host sequence (e.g. a cow milk-specific miRNA that human does not have).

Novel miRNAs and noncoding RNAs

In Table 1, novel miRNA represents the regions that are either annotated as predicted miRNAs in Ensemble [42] or homologs to known miRNAs based on comparison with Rfam sequences and structures using CM model (*e*-value < 1E-5) [35]) and with mature sequences (I_{mature}) >40. Similarly, this pipeline also differentiates other types of noncoding RNA, such as small nucleolar RNAs (snoRNAs), ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), based on the Ensemble genome annotation and similarity with Rfam structure.

Differential expression analysis

Last, for each detected miRNA, we quantified the expression based on its read counts normalized within each sample using

Reads Per Kilobase of transcript per Million mapped reads and across samples using trimmed mean of M values [43]. The differential analysis is performed using EdgeR [44], which outputs visualized results on the expression heatmap, multidimensional scaling, *P*-value and FDR distribution.

Illustration of miRDis using testing sets

The MCF7 breast cancer cell line data (GEO Accession number: GSE31069) has been used to validate the miRNA detection, where four sets of miRNA sequencing library from cytoplasmic fraction and all cell content, before and after Dicer knockdown, respectively, are processed by Illumina Genome Analyzer II [14]. Given that Dicer is an important regulator for miRNA biogenesis, the knockout group is supposed to express less endogenous miRNAs. Here, we compared the performances between our method and existing state-of-the-art tools (Cap-mirseq [15], omiRas [18] and Chimira [24]) on miRNA detection and expression quantification. For the same purpose, we also collected another three sets of sequencing data (Supplementary Table S1) as independent validation on miRNAs from human milk cell (GSE71098), bovine milk exosome (GSE55144) and plant (maize) (GSM1178886-7), respectively.

To illustrate exogenous sequence detection, we use a miRNA sequencing data (SRA ID PRJNA307561) collected from a human milk-feeding study [45]. The miRNA samples were extracted from the blood of five healthy individuals at 0, 3, 6 and 9 h after they consumed 1 L bovine milk. The pooled samples are sequenced using Illumnia-HiSeq2000 at the BGI (Hong Kong, China).

Development of the Web server

The interface of the Web service was developed using smarty/PHP framework while the analytical modules were developed using JavaScript and Ajax, and the data were visualized using JavaScript packages including Datatables and CanvasXexpress. The server is currently hosted in an in-house computer cluster (24 cores and 164G memory) administrated using SGE. To speed up the computation, we parallel all major steps, from uploading and multi-sample processing. Users can access miRDis at http://sbbi.unl.edu/miRDis/index.php.

Results

The schematic flowchart in Figure 2 showcases the implementation workflow and the functionality of this pipeline. Four main components include read processing, read mapping, annotation and differential expression analysis.

Input

MiRDis requires input files as small noncoding RNA sequencing (RNA-seq) read FASTQ data in *.zip or *.gz format. Once uploading all samples, users can input basic parameters or use the default setting on 3′-adapter, minimal and maximal read length and minimum quality in each base. Currently, five host species (human, chimpanzee, dog, rat and mouse) and eight common dietary species including cow, pig, chicken, tomato, maize, soybean, rice and grape are available for the exogenous miRNA analysis. Every job holds a unique job ID (e.g. 20160817135554r); once it is finished, the result can be accessed in miRDis within 2 weeks. An e-mail address is required to receive the job notification from the system.

Output

MiRDis outputs four types of results including summary, identified candidates (lists of annotated known miRNAs, novel miRNAs, other noncoding miRNAs and exogenous miRNAs), annotation (details of each entry) and differential analysis, respectively. For example, Figure 3A shows the result summary from the MCF7 cell line data analysis, where we can see 1212 known miRNA regions (including fragmented regions from the same miRNA), 303 novel miRNAs and 2834 other noncoding miRNAs have been detected. Figure 3B shows the pie charts of read count distribution across all categories, which displays the expression of total miRNA is significantly reduced after Dicer knockout in each group. These pie charts are available for each sample on the summary page.

On the detailed result page (Figure 4A), candidates identified in each category (known, novel, other noncoding and exogenous) are organized by Javascript Datatables, along with the basic information such as mapped region coordinate, sequence similarity and counts in each sample. A further annotation page for each entry shows the details about sequence, structure and genome coordinate, as well as the alignment after mapping. Differential expression analysis is available for any customized group comparisons (Figure 4B).

Case analysis for performance validation

The performance of MiRDis was first evaluated based on the discovery made on the four sets of MCF7 cell line data including 1000 known mature miRNA, 184 known precursors, 303 novel miRNAs and 2834 other noncoding RNAs as summarized in Figure 3A. First, given dicer as an important regulator for miRNA biogenesis, the knockout group is supposed to have reduced expression of miRNAs. Noncoding RNAs (snoRNAs, rRNAs, tRNAs and non-hairpin transcripts) are mostly dicer-independent. Through the expression comparison between control group (SRR326279 and SRR326280) and

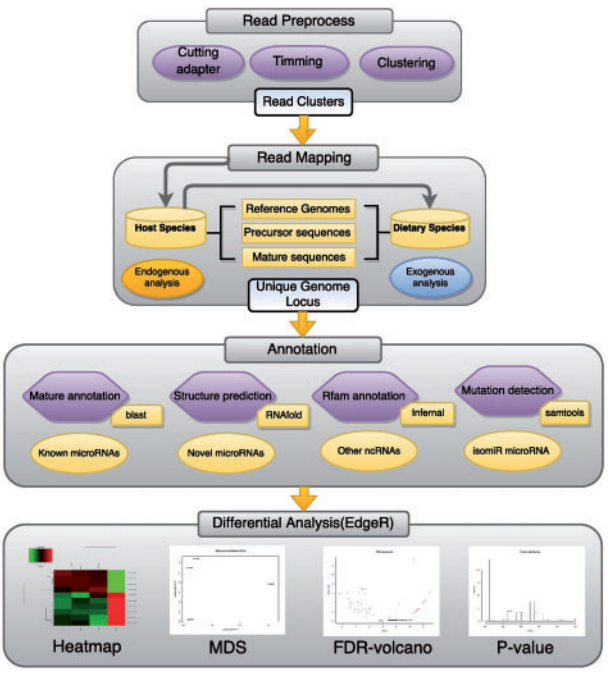

Figure 2. The pipeline workflow for endogenous and exogenous miRNA discovery based on small RNA sequencing analysis.

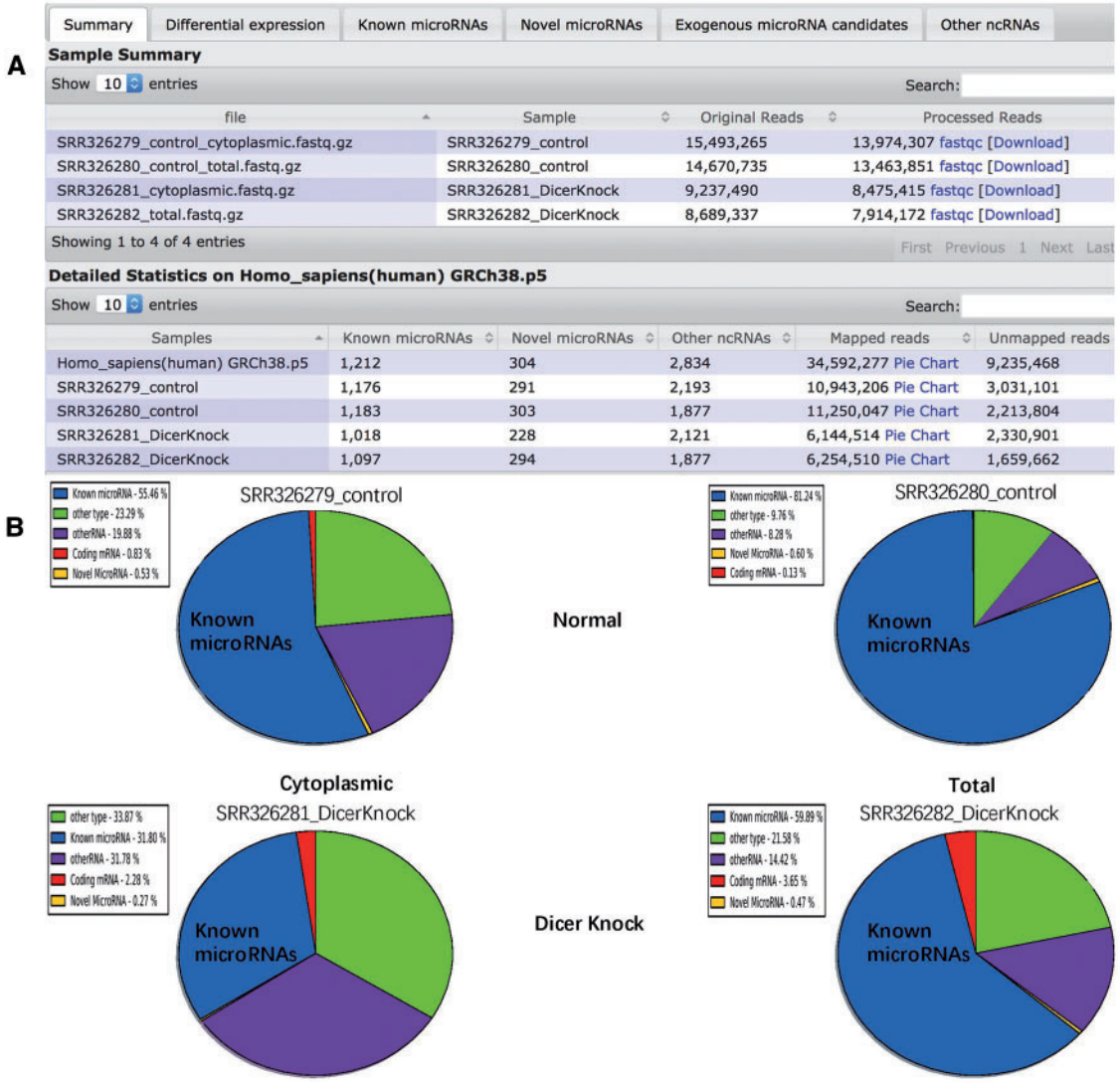

Figure 3. Examples of (**A**) the summary page (from the data analysis on normal and Dicer knockdown MCF7 cells with job ID: 20160817135750r). (**B**) Pie charts that show the proportional abundance among known miRNA, coding mRNA, novel miRNA and other noncoding RNA.

Dicer knockout group (SRR326281 and SRR326282), 559 mapped regions are differentially expressed, which involve 174 known miRNAs and 16 predicted novel miRNAs that are downregulated and 369 noncoding RNAs that are mostly upregulated (Supplementary Table S2).

When compared with other methods such as CAP-miRSeq [15], omiRas [18] and Chimira [24] based on the same analysis, e.g. on sample SRR326279 versus SRR326281, miRDis identified more miRNAs that are suppressed by Dicer silencing, i.e. 676 of 879 compared with 354–477 reported in other methods. We consider this result is better aligned with our understanding of Dicer's role in miRNA biogenesis and shows high sensitivity of miRDis. Similar result holds on the novel miRNA prediction by each tool, e.g. 60.8% in miRDis versus 55.3% in other methods. Note that there is a discrepancy between the detected other noncoding RNAs detected by our tools and others (omiRas), i.e. 544 versus 502. The explanation is that miRDis covers all annotated noncoding RNAs from the Ensemble [42] and Rfam [46], such as tRNA, mitochondrial tRNA, rRNA, small conditional RNA, small nuclear RNA, snoRNA, misc_RNA and long noncoding RNA, while omiRas only focused on the Dicer-independent noncoding RNAs such as PIWI-interacting RNA and snoRNA.

Validation on the other three data sets also shows that miRDis outperforms other methods in terms of the known and novel miRNA detection and computing efficiency (Supplementary Table s1, http://sbbi.unl.edu/miRDis/supp/s1.docx). For instances, in human break milk data, most of other tools detected too many or too few miRNAs (e.g. 737–920 or 35 in average among 20 samples), while miRDis and CAP-miRSeq detected ~500 sequences, which is reasonably close to the current report of ~300 miRNAs confirmed in breast milk cell through microarray and polymerase chain reaction [47, 48]. It is notable that miRDis maintained a high detection rate (70.9–100.0%) while controls the FDR within 11.2% (in contrast to 55.9–74.2 % of other tools). Considering sequencing analysis can render more power in terms of novel sequence detection, this FDR may even be overestimated. Similar observation holds in the bovine milk exosome data where other tools tend to introduce high false positive prediction (>58.2%) compared with the

Figure 4. Output illustration on the MCF7 data, including (**A**) the results on identified candidates in each category, the annotation page for each entry that covers the known sequence and structure from external databases and the alignment details, (**B**) the differential expression analysis on user defined group comparison.

benchmark [49], i.e. 179–307 versus 79. In this regard, miRDis can render up to 92.4% sensitivity by not including any false discovery. As miRDis also offers analysis of plant miRNA in the dietary species, we compared the performance with the tools using public maize sequencing data and found miRDis derived the highest performance (sensitivity = 79.0–92.4%) compared with other tools (sensitivity = 38.4–43.0%).

Regarding the running time, miRDis is among the fastest tools and particularly efficient for uploading (Supplementary Table S1, http://sbbi.unl.edu/miRDis/supp/s1.docx). Via the parallel setting, the current pipeline can handle 40 small RNA-seq data (each includes 50 million reads) and search against nine selected genomes at the same time and finish the job in 12 h.

Case result for exogenous sequence detection

Based on the milk feeding data set, where 67 279 237 reads were mapped to the human genome, we have detected 225 known mature miRNAs, 4 known precursors, 44 novel miRNAs and 839 noncoding RNAs. Among them, 44 miRNAs show elevated abundance after milk feeding (Figure 5).

As introduced in Methods section, all the reads were also mapped to the bovine genome, where we found 68 bovine miRNA candidates that either have specific sequences compared with their human homologs or show elevated expression after milk feeding, although the sequences may be identical. Although the detailed list is provided in Supplementary Table S3, we categorized them into the following four groups (with illustration in Table 3):

(I) miRNAs have specific sequence in bovine compared with human

In this category, mature miRNAs such as miR-181a and miR-655 show slightly different sequences in human and bovine genomes (Table 3_i). According to the read mapping result, both bovine and human miRNAs (bta-miR-181a-2 and has-miR-181a-2) are expressed in the plasma samples (Figure 6A and B). The blue reads in each alignment are those having better alignments in bovine versus human, which supports the expression of exogenous miRNAs.

(II) miRNAs show different sequences in the extended region of bovine and human homologs (Table 2_ii)

Different from Group 1, miRNAs in this category show almost identical mature sequence in bovine and human; however, the differentiation between two different sources can rely on the difference in the extended sequences for each mapped region. For example, hsa-miR-378a-3p and bta-miR-378 share the same annotated mature sequence (ACTGGACTTGGAGTCAGAAGGC) but show different sequences in the extend regions, -(TGG) in bovine and -(CT) in human (Table 3 and Figure 5C and D). This group is less conclusive regarding exogenous sequence detection as the remained adaptor or other factors may cause the additions of the nucleotides in extended.

(III) miRNAs that are specific in bovine and has no annotation in human

There are two cases that bovine has specific mature sequence, while human does not have (bta-miR-2898 and -miR-1839). In Table 3, the exogenous potential of bta-miR-2898 is 2.33, which means over 33% mapped reads have better mapped in bovine genome. Bta-miR-1839 (bovine 14:1883889-1883907) shows lower potential score as 2, as a similar sequence region

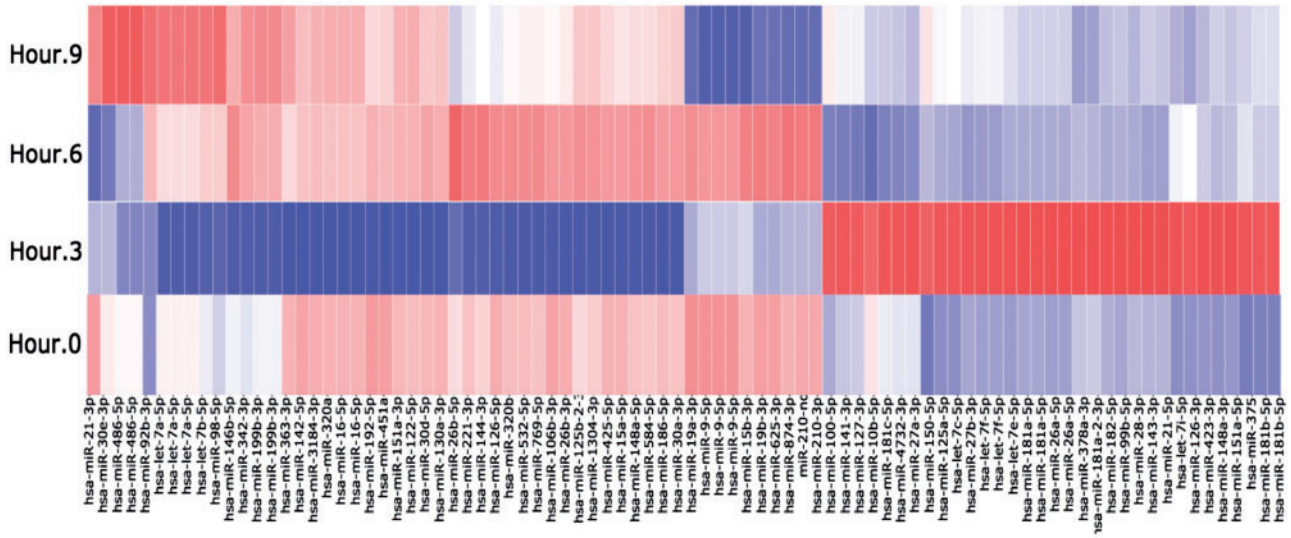

Figure 5. Heatmap shows the altered expression of 58 miRNAs (44 upregulated) in human plasma after milk feeding at 3, 6 and 9 h (with the job ID: 20160817135554r). The expression is scaled within [-2, 2].

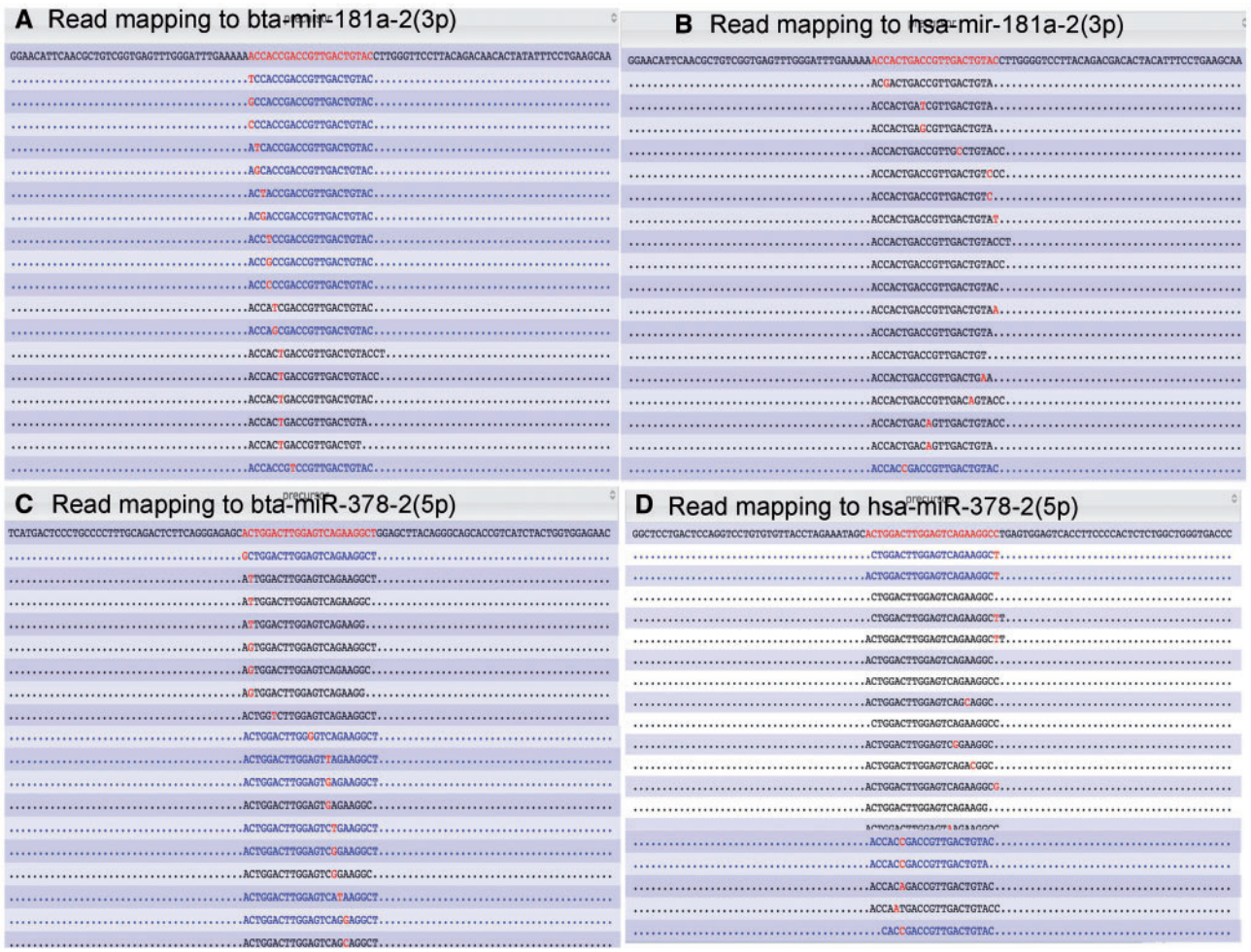

Figure 6. Read alignments in both bovine and human miRNA regions (**A** and **B**) miR181a and (**C** and **D**) miR-378. Reads in black represent those have better alignments in human versus bovine, while reads in blue represent the contrary.

Table 2. Performance of miRDis compared with Cap-mirSeq, omiRAS and chimera

Types	miRDis	CAP-miRSeq	omiRAS	Chimira
Version	v1.0	v1.1	12/2013	V1.0
Aligner	Bowtie	Bowtie	Bowtie	BLASTN
Reference genome	hg19	hg19	hg19	hg19
Computation time for single data set (219–387 MB)	2 h	5 h	3 h	1 h
Downregulated/mature miRNAs	676/879 (76.9%)	354/769 (46.0%)	362/692 (52.3%)	477/913 (52.2%)
Downregulate/novel miRNAs	185/303 (60.8%)	183/353 (51.8%)	145/262 (55.3%)	–
Differentially expressed noncoding RNAs	397/544 (72.9%)	–	403/502 (80.3%)	–

Note. The common settings include the extraction of mapped region with more than five supporting reads and two mismatches for mapping.

Table 3. Different categories of miRNAs that show exogenous potential at different level based on the sequence and expression evidence (Group 5 is listed in Supplementary Table S3)

Categories	Mapped regions in bovine	Annotation type	Sequences (bovine versus human)	Exogenous potential	Altered expression Log_2FC
I	11:95709486–95709507	Known precursor	bta-miR-181* ACCACCGACCGTTGACTGTAC has-miR-181a-3p ACCATCGACCGTTGACTGTACC	0.98	3.39
	12:19596253–19596275	Known mature	bta-miR-16a TAGCAGCACGTAAATATTGGTG hsa-miR-16b-5p TAGCAGCACGTAAATATTGGCG	0.8	–
	21:67587374–67587396	Known mature	bta-miR-655 ATAATACATGGTTAACCTCTCT hsa-miR-655-3p ATAATACATGGTTAACCTCTTT	0.66	–
II	4:10715304–10715327	isoMiR-mature	bta-miR-378 ACTGGACTTGGAGTCAGAAGGC(TGG) hsa-miR-378a-3p ACTGGACTTGGAGTCAGAAGGC(CT)	0.79	0.72
	18:56407899-56407922	Known mature	bta-miR-150 TCTCCCAACCCTTGTACCAGTGT(GT) has-miR-150 TCTCCCAACCCTTGTACCAGTG(CT)	0.1	2.35
	7:62809358–62809378	Known mature	bta-miR-143 TGAGATGAAGCACTGTAGCTCG hsa-miR-143 TGAGATGAAGCACTGTAGCTC(A)	0.00039	1.12
	27:36261887–36261910	isoMiR-mature	bta-miR-486 TCCTGTACTGAGCTGCCCCGAG(GC) hsa-miR-486(mir-486-2) TCCTGUACTGAGCTGCCCCGAG(CU)	2.71E-06	1.5
III	8:74354026–74354044	Known mature	bta-miR-2898 TGGTGGAGATGCCGGGGA hsa NA	2.33	−0.7
	14:1883889–1883907	Known mature	bta-miR-1839 AAGGTAGATAGAACAGGTCTTGTT hsa NA	2	−0.84

in human (human 15:82756012-82756031) is annotated as other noncoding RNA (SCARNA15: RF00426) according to infernal prediction with *e*-value = 1.3E-21. However, both cases show decreased abundance after milk intake, which makes them less compelling as exogenous miRNA candidates.

(IV) miRNAs that have identical sequences in human and bovine but show elevated expression after milk feeding

In total, 42 miRNAs are included in this group (Supplementary Table S3). We consider entries in this category the least conclusive toward exogenous sequence identification. The differentiation of identical sequences from human and bovine requires novel laboratory protocols.

According to both sequence and expression evidence, we proposed that three bovine miRNAs (including bta-mir-181a-2, -miR-378 and -miR-150) are possible exogenous miRNAs, plus two additional (bta-miR-143 and -miR-486) weaker cases. It is also interesting to find in literatures that miR-378 regulates fatty acid and cholesterol metabolism pathways by targeting lipid metabolism genes related to milk fat metabolism in bovine primary mammary epithelial cells [50, 51]. Similarly, bta-miR-181a regulates the biosynthesis of bovine milk fat by targeting ACSL1 [52], while miR-150 inhibits the synthesis of the transcription factor c-Myb to regulate B-cell differentiation [53]. It is hypothetical that they may play similar roles in human system.

Discussion

In this work, we present miRDis, a Web-based small RNA sequencing analytical pipeline that displays the following key features (i) systematic annotation of known miRNAs and other noncoding RNAs based on read mapped regions, (ii) prediction of novel miRNAs and noncoding RNAs through assigning ambiguous reads to unique genome region with well-predicted RNA structure, (iii) detection of candidate exogenous miRNAs transported from dietary species and (iv) support of the comparative differential expression analysis. Through a simple graphical interface, users can use the full analysis of this one-stop tool for miRNA sequencing data analysis through minimal parameter settings. The tabular and graphical output contains detailed reports on the read alignment, annotation and other related statistics.

MiRDis has been tested on small noncoding RNA sequencing data from human milk-feeding study, where we detected a few candidate exogenous miRNAs in human plasma based on both sequence and expression evidence. Through the data visualization, users can examine the detailed alignment associated with both exogenous and endogenous cases. To validate the expression quantification, we used the Dicer silencing data set as benchmark to compare miRDis with other existing tools. In general, miRDis can identify more miRNAs than others tools

because of the integrated annotation from both miRBase and Ensemble database. Both known and predicted miRNAs shows consistently lower expression with Dicer knockdown treatment, while other noncoding miRNAs such as snoRNAs, rRNAs, tRNAs and non-hairpin transcripts are independent on Dicer regulation. Other existing methods can be applied for functional analysis of the identified miRNAs, either experimentally or computationally [54–56].

Note that the capacity of this pipeline will be significantly advanced by compressing reads with the same sequence from different samples into a unique sequence set. In this way, miRDis annotates consensus sequence of the mapped regions using all reads across multiple samples instead of from a single sample to ensure a higher recall rate while users can easily retrieve every sample count by information imbedded in the read tags. It significantly increases the performance in terms of computation time on mapping and memory consumption. Based on a new test on large-scale cancer miRNA-seq data from The Cancer Genome Atlas, the current pipeline can handle 40 small RNA-seq data (each includes 40 million reads) and search against nine selected genomes (one host and eight dietary species) at the same time and finish the job in 12 h.

Conclusions

In summary, we proposed the first pipeline for small noncoding RNA sequencing data analysis that enables the automated detection in host samples the presence of both endogenous miRNA and exogenous miRNA from dietary species. In addition, the improvement in performance of our system over state-of-the-art methods lies in the high sensitivity of miRNA detection and expression quantification, and the differentiation of the isomiRs and non-host miRNAs. We also overcome the challenges of scaling this system for processing large set of miRNA-seq data through parallel computation. With increased research efforts in miRNA biology, we believe miRDis provides an efficient and friendly tool for making promising discoveries in miRNA cross-species transport.

Key Points

- Compelling evidence shows that animals can acquire exogenous miRNA from diet; however, the mechanism of cross-species transport miRNA has yet to be fully explored.
- It is now possible to identify in host species the exogenous miRNA sequences using computational analysis of small noncoding RNA sequencing data.
- Computational methods for the miRNA sequencing analysis have been recently developed, and none has focused on the automatic detection of the exogenous sequences yet.
- We have developed a new system for the comprehensive discovery of miRNA of all kinds based on the sequencing data analysis, which represents the first automated tool for exogenous miRNA detection.
- We used several public benchmark noncoding miRNA data sets and an in-house data from a human feeding study to compare the performance of the state-of-the art methods for miRNA detection and provide the results.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The authors would like to thank all SBBI members who have been involved in this work for providing helpful discussions and technical assistance.

Funding

The National Institutes of Health funded COBRE grant (grant number 1P20GM104320), UNL Food For Health seed grant and the Tobacco Settlement Fund as part of the Cui's start-up grant support.

References

1. Dong H, Lei J, Ding L, *et al*. MicroRNA: function, detection, and bioanalysis. *Chem Rev* 2013;**113**(8):6207–33.
2. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004;**116**(2):281–97.
3. Zhang L, Hou D, Chen X, *et al*. Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Res* 2012;**22**(1):107–26.
4. Wang K, Li H, Yuan Y, *et al*. The complex exogenous RNA spectra in human plasma: an interface with human gut biota? *PLoS One* 2012;**7**(12):e51009.
5. Baier SR, Nguyen C, Xie F, *et al*. MicroRNAs are absorbed in biologically meaningful amounts from nutritionally relevant doses of cow milk and affect gene expression in peripheral blood mononuclear cells, HEK-293 kidney cell cultures, and mouse livers. *J Nutr* 2014;**144**(10):1495–500.
6. Zhou Z, Li X, Liu J, *et al*. Honeysuckle-encoded atypical microRNA2911 directly targets influenza A viruses. *Cell Res* 2015;**25**(1):39–49.
7. Baier SR, Nguyen C, Xie F, *et al*. MicroRNA from cow's milk are bioavailable and affect gene expression in humans. *J Nutr* 2014, in press.
8. Izumi H, Kosaka N, Shimizu T, *et al*. Bovine milk contains microRNA and messenger RNA that are stable under degradative conditions. *J Dairy Sci* 2012;**95**(9):4831–41.
9. Arnold CN, Pirie E, Dosenovic P, *et al*. A forward genetic screen reveals roles for Nfkbid, Zeb1, and Ruvbl2 in humoral immunity. *Proc Natl Acad Sci USA* 2012;**109**(31):12286–93.
10. Liu R, Ma X, Xu L, *et al*. Differential microRNA expression in peripheral blood mononuclear cells from Graves' disease patients. *J Clin Endocrinol Metab* 2012;**97**(6):E968–72.
11. Etheridge A, Gomes CP, Pereira RW, *et al*. The complexity, function and applications of RNA in circulation. *Front Genet* 2013;**4**:115.
12. Kitchen R, Subramanian SL, Navarro F, *et al*. A comprehensive method for the analysis of extracellular small RNA-seq data, including characterisation based on cellular expression profiles and exogenous sequence detection (Abstract). The Fourth International Meeting of ISEV, ISEV2015. *J Extracell Vesicles* 2015;**4**:27783.
13. Lukasik A, Zielenkiewicz P. In silico identification of plant miRNAs in mammalian breast milk exosomes–a small step forward? *PLoS One* 2014;**9**(6):e99963.

14. Friedlander 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.
15. Sun Z, Evans J, Bhagwate A, *et al*. CAP-miRSeq: a comprehensive analysis pipeline for microRNA sequencing data. *BMC Genomics* 2014;**15**:423.
16. Huang PJ, Liu YC, Lee CC, *et al*. DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res* 2010;**38**: W385–91.
17. 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.
18. Muller 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.
19. Barturen G, Rueda A, Hamberg M, *et al*. sRNAbench: profiling of small RNAs and its sequence variants in single or multispecies high-throughput experiments. *Methods Next Gener Seq* 2014;**1**:21–31.
20. Axtell MJ. ShortStack: comprehensive annotation and quantification of small RNA genes. *RNA* 2013;**19**(6):740–51.
21. Yang X, Li L. miRDeep-P: a computational tool for analyzing the microRNA transcriptome in plants. *Bioinformatics* 2011;**27**(18):2614–15.
22. Lei J, Sun Y. miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. *Bioinformatics* 2014;**30**(19):2837–9.
23. Rueda A, Barturen G, Lebron R, *et al*. sRNAtoolbox: an integrated collection of small RNA research tools. *Nucleic Acids Res* 2015;**43**(W1):W467–73.
24. Vitsios DM, Enright AJ. Chimira: analysis of small RNA sequencing data and microRNA modifications. *Bioinformatics* 2015;**31**(20):3365–7.
25. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;**42**:D68–73.
26. Pruitt KD, Brown GR, Hiatt SM, *et al*. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* 2014;**42**: D756–63.
27. Nawrocki EP, Burge SW, Bateman A, *et al*. Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res* 2015;**43**:D130–7.
28. Wang X, Liu XS. Systematic curation of miRBase annotation using integrated small RNA high-throughput sequencing data for *C. elegans* and Drosophila. *Front Genet* 2011;**2**:25.
29. Ameres SL, Zamore PD. Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol* 2013;**14**(8):475–88.
30. Vickers KC, Sethupathy P, Baran-Gale J, *et al*. Complexity of microRNA function and the role of isomiRs in lipid homeostasis. *J Lipid Res* 2013;**54**(5):1182–91.
31. Wyman SK, Knouf EC, Parkin RK, *et al*. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. *Genome Res* 2011;**21**(9):1450–61.
32. Smalheiser NR, Torvik VI. Mammalian microRNAs derived from genomic repeats. *Trends Genet* 2005;**21**(6):322–6.
33. Berezikov E. Evolution of microRNA diversity and regulation in animals. *Nat Rev Genet* 2011;**12**(12):846–60.
34. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, *et al*. ViennaRNA package 2.0. *Algorithms Mol Biol* 2011;**6**:26.
35. Nawrocki EP, Eddy SR. Computational identification of functional RNA homologs in metagenomic data. *RNA Biol* 2013;**10**(7):1170–9.
36. Bioinformatics B. *FastQC A Quality Control Tool for High Throughput Sequence Data*. Cambridge, UK: Babraham Institute 2011.
37. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J* 2011;**17**(1):10–12.
38. Tonge DP, Gant TW. What is normal? Next generation sequencing-driven analysis of the human circulating miRNAOme. *BMC Mol Biol* 2016;**17**:4.
39. Kersey PJ, Allen JE, Christensen M, *et al*. Ensembl genomes 2013: scaling up access to genome-wide data. *Nucleic Acids Res* 2014;**42**:D546–52.
40. Langmead B, Trapnell C, Pop M, *et al*. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* 2009;**10**(3):R25.
41. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. *Curr Protoc Bioinformatics* 2014;**47**:11.12.1–34.
42. Guttman M, Amit I, Garber M, *et al*. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 2009;**458**(7235):223–7.
43. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;**11**(3):R25.
44. 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.
45. Shu J, Chiang K, Zempleni J, *et al*. Computational characterization of exogenous MicroRNAs that can be transferred into human circulation. *PLoS One* 2015;**10**(11):e0140587.
46. Griffiths-Jones S, Moxon S, Marshall M, *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 2005;**33**:D121–4.
47. Munch EM, Harris RA, Mohammad M, *et al*. Transcriptome profiling of microRNA by next-gen deep sequencing reveals known and novel miRNA species in the lipid fraction of human breast milk. *PLoS One* 2013;**8**(2):e50564.
48. Hassiotou F, Alsaweed M, Savigni D, *et al*. Profiling of human milk miRNA. *FASEB J* 2015;**29(1 Suppl)**:582.8.
49. Izumi H, Tsuda M, Sato Y, *et al*. Bovine milk exosomes contain microRNA and mRNA and are taken up by human macrophages. *J Dairy Sci* 2015;**98**(5):2920–33.
50. Shen B, Zhang L, Lian C, *et al*. Deep sequencing and screening of differentially expressed MicroRNAs related to milk fat metabolism in Bovine primary mammary epithelial cells. *Int J Mol Sci* 2016;**17**(2): 200.
51. Gerin I, Bommer GT, McCoin CS, *et al*. Roles for miRNA-378/378* in adipocyte gene expression and lipogenesis. *Am J Physiol Endocrinol Metab* 2010;**299**(2):E198–206.
52. Lian S, Guo JR, Nan XM, *et al*. MicroRNA Bta-miR-181a regulates the biosynthesis of bovine milk fat by targeting ACSL1. *J Dairy Sci* 2016;**99**(5):3916–24.
53. Xiao C, Calado DP, Galler G, *et al*. MiR-150 controls B cell differentiation by targeting the transcription factor c-Myb. *Cell* 2007;**131**(1):146–59.
54. Zeng X, Zhang X, Zou Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief Bioinform* 2016;**17**(2):193–203.
55. Wong N, Wang X. miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res* 2015;**43**:D146–52.
56. Mullokandov G, Baccarini A, Ruzo A, *et al*. High-throughput assessment of microRNA activity and function using microRNA sensor and decoy libraries. *Nat Methods* 2012;**9**(8):840–6.