Published online 4 December 2012 *Nucleic Acids Research, 2013, Vol. 41, No. 2* **727–737**
doi:10.1093/nar/gks1187

miRDeep*: an integrated application tool for miRNA identification from RNA sequencing data

Jiyuan An[1,*], John Lai[1], Melanie L. Lehman[1,2] and Colleen C. Nelson[1,2,*]

[1]Australian Prostate Cancer Research Centre—Queensland, Institute of Health and Biomedical Innovation (IHBI), Queensland University of Technology, Princess Alexandra Hospital, Level 1, Building 1, Ipswich Road, Brisbane, Queensland, QLD 4102, Australia and [2]Department of Urologic Sciences, Prostate Center, Vancouver General Hospital, University of British Columbia, Vancouver, British Columbia, V5Z 1M9, Canada

Received April 22, 2012; Revised October 27, 2012; Accepted October 29, 2012

ABSTRACT

miRDeep and its varieties are widely used to quantify known and novel micro RNA (miRNA) from small RNA sequencing (RNAseq). This article describes miRDeep*, our integrated miRNA identification tool, which is modeled off miRDeep, but the precision of detecting novel miRNAs is improved by introducing new strategies to identify precursor miRNAs. miRDeep* has a user-friendly graphic interface and accepts raw data in FastQ and Sequence Alignment Map (SAM) or the binary equivalent (BAM) format. Known and novel miRNA expression levels, as measured by the number of reads, are displayed in an interface, which shows each RNAseq read relative to the pre-miRNA hairpin. The secondary pre-miRNA structure and read locations for each predicted miRNA are shown and kept in a separate figure file. Moreover, the target genes of known and novel miRNAs are predicted using the TargetScan algorithm, and the targets are ranked according to the confidence score. miRDeep* is an integrated standalone application where sequence alignment, pre-miRNA secondary structure calculation and graphical display are purely Java coded. This application tool can be executed using a normal personal computer with 1.5 GB of memory. Further, we show that miRDeep* outperformed existing miRNA prediction tools using our LNCaP and other small RNAseq datasets. miRDeep* is freely available online at http://www.australianprostatecentre.org/research/software/mirdeep-star.

INTRODUCTION

Micro RNA (miRNA) is a class of small, non-protein–coding RNA (ncRNA) that is important in normal physiology, which includes development and tissue-specific processes in many eukaryotic systems. Mature miRNAs are typically generated from longer primary and precursor miRNA or from intronic RNA (1, 2). miRNA typically mediate its biological effects through translation inhibition or, in some instances, by RNA degradation through the RNA-induced silencing complex (RISC) (3). It is thought that, similar to other diseases, dysregulated miRNA expression in prostate cells can lead to prostate cancer progression. Indeed, 26 miRNAs have been found to be deregulated in prostate cancer (4). The prostate is regulated by the male hormones, androgens, and the action of androgens is mediated by its cognate receptor, the androgen receptor (AR), which is a ligand-dependent transcription factor. Concomitantly, androgens are also important in prostate cancer progression (5). Consequently, much research in the prostate cancer field has focused on genes that are targeted by the AR signaling axis in this disease. However, other than the TMPRSS2–ERG fusion gene (6), which appears to be overexpressed in many prostate cancers (7), the other bona fide AR target genes that are important in prostate cancer progression remain elusive. Two recent microarray studies suggest that at least 27 known miRNAs are androgen regulated in prostate cancer cells (8,9), although this number is likely to increase as data emerge from next-generation sequencing platforms that have already identified many novel prostate expressed miRNA (10–12).

The advent of high-throughput sequencing technology has provided researchers an unbiased opportunity to systematically identify most, if not all, of the miRNA that are expressed in the transcriptome. Thus, determining levels

*To whom correspondence should be addressed. Tel: +61 7 3176 3075; Fax: +61 7 3176 7445; Email: j.an@qut.edu.au
Correspondence may also be addressed to Colleen C. Nelson. Tel: +61 7 3176 7443; Fax: +61 7 3176 7440; Email: colleen.nelson@qut.edu.au

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

of known and novel miRNA from small RNA sequencing (RNAseq) data is an important issue in the era of next generation sequencing. Although there are several miRNA profiling applications such as miRanalyzer (13), miRTRAP (14) and MIReNA (15), these methods rely on known miRNAs and a user's training data. Consequently, these results largely depend on known miRNA data and the classification algorithm. In miRDeep (16), prediction of miRNA from sequenced reads is output as a probability from the log odds ratio. More recently, miRDeep2 was developed (17), which offers similar improved mature miRNA prediction algorithms as those used in our miRDeep* tool. Indeed, miRDeep2 has shown great predictive ability of bona fide miRNA by comparing between datasets where the miRNA biogenesis pathway was or was not active. miRDeep2 has also adopted other improved functionalities such as a graphics output of the predicted secondary structure of pre-miRNA. However, miRDeep2 and other similar miRNA prediction tools are dependent on other software such as pre-miRNA secondary structure prediction and/or genome mapping. Consequently, we have developed miRDeep*, which is an integrated tool that can be used to identify novel miRNA from raw RNAseq reads, as well as quantifying miRNA expression. Further, miRDeep* offers a user-friendly graphical output that displays the location of the sequenced reads in the pre-miRNA hairpin structure. All components of miRNA identification in this application, such as sequence alignment (18, 19) and RNA folding (20), is purely Java coded. In addition, miRDeep* incorporates the widely used TargetScan program (21–23). The targets of both known and novel miRNA are predicted using one click. Finally, miRDeep* is compatible for use with other species to identify both novel and known miRNAs from raw small RNAseq data.

In this study, we have used the Illumina Genome Analyser II to sequence the androgen-regulated small RNA transcriptome of LNCaP prostate cancer cells. We demonstrate using this dataset that miRDeep* has higher precision in detecting novel miRNA from the total number of predictions when compared with other available miRNA prediction tools, such as miRDeep, miRDeep2, miRanalyzer, miRTRAP and MIReNA.

MATERIALS AND METHODS

Cell culture

The LNCaP prostate cancer cell line was obtained from American Type Culture Collection. Cells were maintained in RPMI 1640 media (Life Technologies, Carlsbad, CA, USA) supplemented with 10% fetal calf serum (Life Technologies). For steroid treatment experiments, cells were cultured until approximately 70% confluent and then maintained in media containing 10% charcoal-stripped serum (HyClone, Denver, CO) for 48 h, followed by the addition of 1 nM R1881 or ethanol (Mock control) for an additional 48 h.

Small RNAseq

RNA was isolated using the miRVana RNA isolation kit (Life Technologies). The RNA was prepared for sequencing using the Illumina Small RNA Sample Prep Kit. The small RNA libraries were sequenced on the Illumina Genome Analyzer II at the Genome Sciences Centre (Vancouver, BC) as described (24).

miRDeep* strategy for prediction of miRNA

The miRDeep* pipeline for the prediction of miRNA from small RNAseq data is summarized in Figure 1a. Genomic mapping is one of the most time-consuming procedures. To reduce mapping time, all sequence reads that are the same are first aggregated, and representative reads are then mapped to the genome. Moreover, reads that are outside the range of miRNA length (18–23 nucleotides) are excluded before genomic mapping. Mapped reads are then output in a SAM format (25). Alternatively, pre-generated SAM files by other genomic tools can be used for integration into the miRDeep* pipeline.

The miRDeep (16) algorithm aggregates reads that align to the same strand within 30 bp of each other to form a read coverage region. However, miRDeep* adopts a different strategy to excise the potential precursor locus range. In miRDeep, if the coverage region is greater than 30 bp, then both sides of the coverage region are extended by 22 bp for the precursor miRNA (Figure 1b; Case 1). However, if the coverage region is less than 30 bp, then one side of the coverage is extended by 22 bp and the other side is extended to ensure that the potential miRNA precursor totals 110 bp in length (Figure 1b; Case 2). The highest expressed read is considered as the mature miRNA.

In miRDeep*, the highest expressed read at the potential miRNA locus is also considered as the mature miRNA. One side of the read is extended by 22 bp (for miRNA offset RNA), and the other side of the read is extended by 15 bp (to account for the loop region) and the mature RNA* that has the same number (*n* [base pairs]) of nucleotides as the mature miRNA, as well as by 22 bp for the miRNA offset RNA (Figure 1d). Consequently, miRDeep* can prevent some false negatives that might arise from improper precursor regions that can be selected for miRDeep. This strategy is similar to that of the recently published miRDeep2 algorithm (Figure 1c).

miRDeep* application and interface

This application tool has been developed in the platform-independent computer language, Java. As a result, it can run in Linux, MS Windows and MacOS. To further reduce the memory requirement for miRDeep*, we have implemented a compact Java version of the Bowtie alignment algorithm (18) where the index files are loaded chromosome by chromosome, instead of loading the entire chromosome index at once.

Using this approach, it takes about 1 h to align 3 million reads using a PC comprising 2.5 GHz CPU and 4.00 GB memory (RAM). From the raw data (FastQ format),

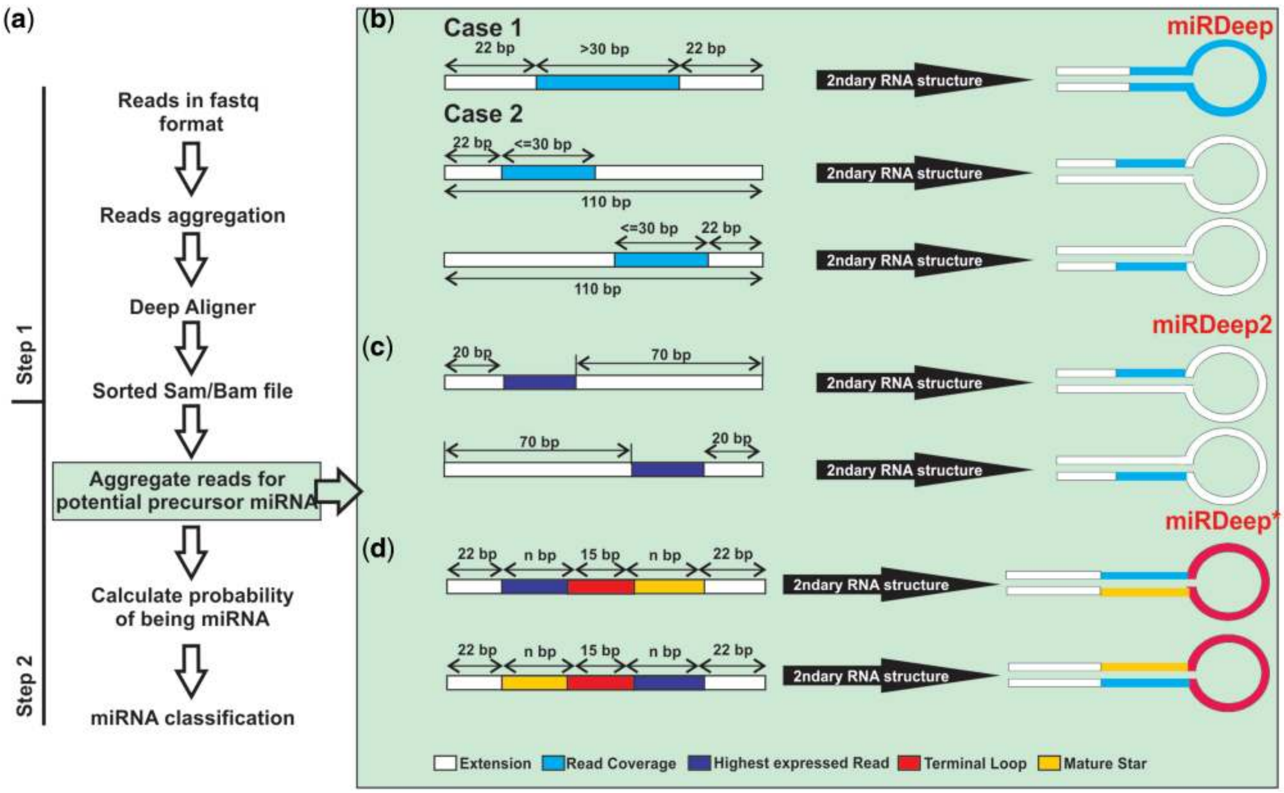


Figure 1. miRNA prediction strategy for miRDeep, miRDeep2 and miRDeep*. (**a**) miRNA prediction pipeline for miRDeep*. Step 1 involves mapping sequences to the genome to produce a SAM file. In step 2, aggregated RNAseq reads are assessed for pre-miRNA secondary structure potential. (**b**) The two different miRNA prediction strategies (case 1 and case 2) for miRDeep. (**c**) and (**d**) Modified miRNA prediction strategy for miRDeep2 and miRDeep*.

miRDeep* will identify novel miRNAs and also generate all expression data for known and novel miRNAs.

The miRDeep* interface has four parts:

(1) User input parameters: The Illumina Deep sequencing platform adapters are the default adapter sequence for miRDeep*. The default length of miRNAs is set at 18–23 nucleotides. Low-quality reads are filtered out at the alignment stage. A read with more than a 20 phred score (which represents 99% base call accuracy) is considered to be a good read. We also filter out multi-mapping reads with alignments to more than 100 genomic loci. This parameter can be changed at the user's discretion. miRDeep* uses other mapped reads at that loci to determine the strength of the prediction (Figure 2). Reads mapping to tRNA, sno-RNA and piRNA are also excluded. Currently, miRDeep* is unable to handle technical or biological mismatches. The score is based on the probabilistic score of the potential miRNA precursor that was previously described (26, 27). miRDeep* also incorporates a tick box for to allow miRDeep* users to alternatively use the original miRDeep algorithm.

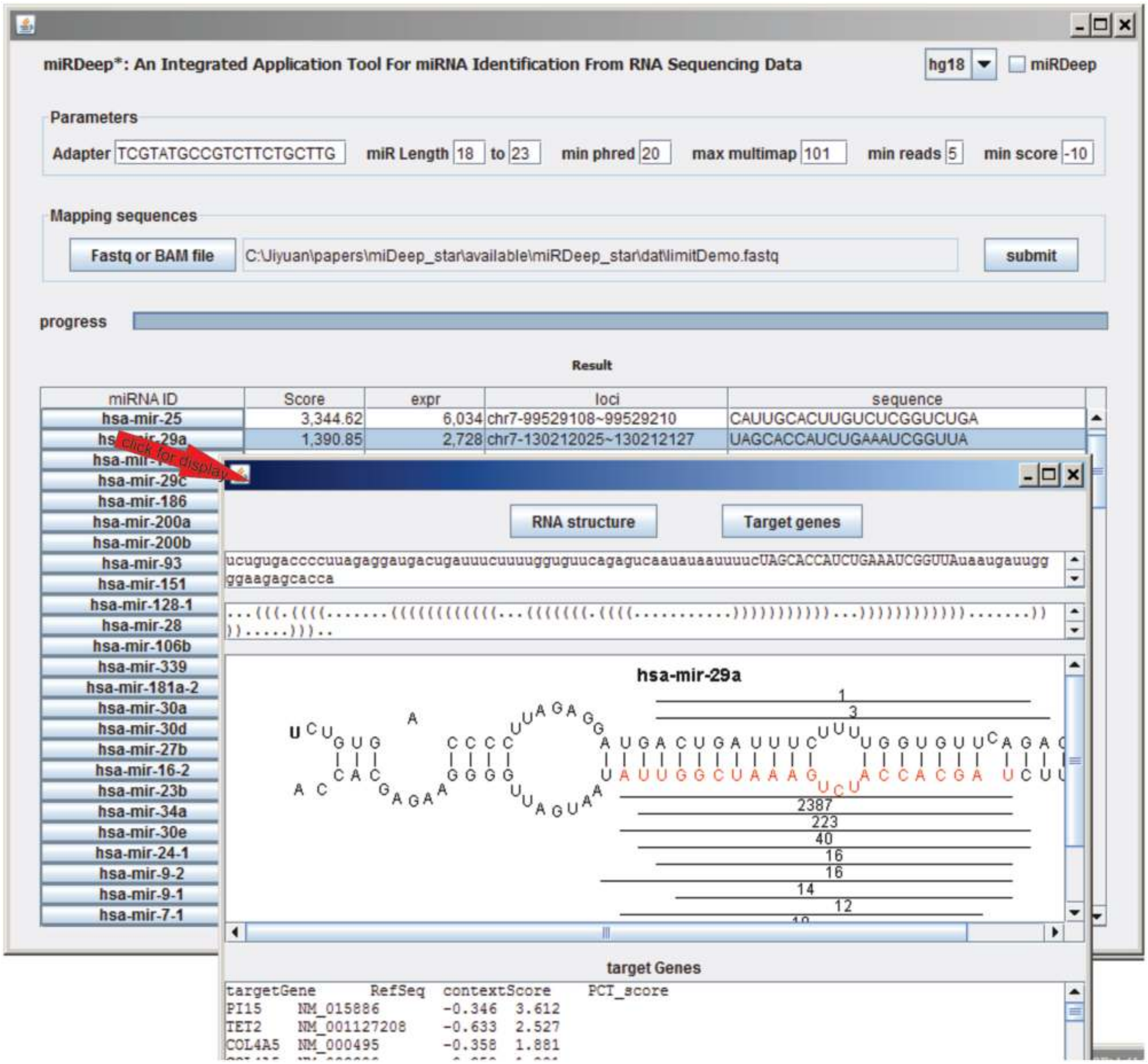
(2) Input file area for raw data (FastQ), pre-aligned files (SAM or BAM) or results (generated by miRDeep*) files: In the first step, input reads are converted to FastQ format (25) for genomic alignment. We have implemented our small memory consuming Bowtie algorithm (18) for the alignments because miRDeep* is targeted at ordinary PCs with limited memory capacities. Users can also employ alternative alignment tools, such as soap2 (19). The pre-aligned BAM/SAM file is inputted into the textbox.

(3) A progress bar: To monitor the progress of miRDeep*, the percentage of completion of genomic mapping and identifying miRNA are displayed at the all procedure stages.

(4) Results display: There are five output files located in the same directory as the input data file. These files are as follows:

(i) xxx.result: Contains information such as chromosomal location, expression levels, pre-miRNA sequences, secondary structure of pre-miRNA and mature miRNA sequences for miRDeep*-predicted miRNA. The miRNA ID for predicted miRNA are assigned to known miRNA if the mature miRNA sequences match. Pre-miRNA secondary structure prediction is rewritten in Java based on the Vienna algorithm (28). All default parameters are used for prediction, but in the output figure, the RNA secondary structure is shown with G–U pairing.

(ii) xxx.cluster: Contains information for clusters of predicted miRNA. All mature miRNA sequences in a cluster are from the same reads. In the file,



miRDeep*: An Integrated Application Tool For miRNA Identification From RNA Sequencing Data   hg18 ▼ □ miRDeep

Parameters

Adapter TCGTATGCCGTCTTCTGCTTG miR Length 18 to 23 min phred 20 max multimap 101 min reads 5 min score -10

Mapping sequences

Fastq or BAM file C:\Uiyuan\papers\miDeep_star\available\miRDeep_star\dat\limitDemo.fastq submit

progress

Result

miRNA ID	Score	expr	loci	sequence
hsa-mir-25	3,344.62	6,034	chr7-99529108~99529210	CAUUGCACUUGUCUCGGUCUGA
hsa-mir-29a	1,390.85	2,728	chr7-130212025~130212127	UAGCACCAUCUGAAAUCGGUUA
hsa-mi				
hsa-mir-29c				
hsa-mir-186				
hsa-mir-200a				
hsa-mir-200b				
hsa-mir-93				
hsa-mir-151				
hsa-mir-128-1				
hsa-mir-28				
hsa-mir-106b				
hsa-mir-339				
hsa-mir-181a-2				
hsa-mir-30a				
hsa-mir-30d				
hsa-mir-27b				
hsa-mir-16-2				
hsa-mir-23b				
hsa-mir-34a				
hsa-mir-30e				
hsa-mir-24-1				
hsa-mir-9-2				
hsa-mir-9-1				
hsa-mir-7-1				

click for display

RNA structure Target genes

```
ucugugaccccuuagaggaugacugauuucuuuugguguucagagucaauauaauuuucUAGCACCAUCUGAAAUCGGUUAuaaugauugg
ggaagagcacca
```

```
...(((.((((.......((((((((((((...(((((((.((((...........)))))))))))...)))))))))))).......))
)).....)))..
```

target Genes

```
targetGene      RefSeq   contextScore    PCT_score
PI15    NM_015886        -0.346  3.612
TET2    NM_001127208     -0.633  2.527
COL4A5  NM_000495        -0.358  1.881
```

Figure 2. The input/output interface for miRDeep*. Parameters are customizable for the genome build, use of the original mirDeep algorithm (miRDeep), filtering out platform-specific adapter sequences (Adapter), the miRNA length (miR Length), the minimum read quality (min phred), the maximum number of loci mapped for a particular read (max multimap), the minimum number of reads (min reads) and the minimum prediction score (min score). The location of the mature miRNA and the number of reads relative to the pre-miRNA secondary structure can also be displayed by simply clicking on the miRNA of choice. On the pop-up window, target genes can be predicted for current sequence by clicking 'Target genes'. The sequence in the top textbox can be replaced and click 'RNA structure' to regenerate RNA secondary structure. Note that mature sequence needs to be in upper case.

the first column refers to miRNAs that share the same sequence reads, and the second column refers to the number of shared reads.

(iii) xxx.sam: All reads in the FastQ file are mapped to a coordinate sorted SAM file. A new attribute, called XS, is created, which records the copy number of a sequenced read.

(iv) xxx.knownMiR: Contains two columns, one refers to the miRNA ID and the other refers to the number of sequenced reads mapping to the genomic region where the miRNA is located.

(v) xxx.purified.fa: Contains purified reads in FastQ format based on three types of filtering. (i) Reads that do not conform to mature miRNA length (18–23 bp). (ii) Ambiguous nucleotides 'N' appear in the reads. (iii) The quality of a read is lower than the threshold. To reduce mapping time, reads that have the exact sequence are aggregated and the number is put in the second item of the description line of the fasta format.

The expression level of each predicted miRNA is determined by the number of reads covering the mature miRNA region. miRDeep* accounts for miRNA isoforms by including variants that share 90% sequence homology with the mature miRNA.

miRDeep* output visualization

miRDeep*-predicted miRNA is shown in a pop-up window by clicking the miR ID in the first column of the output table (Figure 2). The actual sequence reads are also displayed in the pre-miRNA (Figure 2). The mature miRNA sequence is shown in red text. Users can also change the sequence in the top textbox to see the pre-miRNA secondary structure by clicking the 'RNA structure' button in the pop-up window. The targets can be obtained by clicking 'Target genes'. The details are given below.

Prediction of target genes

There are many miRNA target prediction tools, such as Diana microT (29), miRanda (30), TargetMiner (31) and MirTarget (32). In this application, we use one of the most commonly used prediction tools, TargetScan, (21) to predict targets for identified known and novel miRNAs. The list of target genes is ranked in ascending order of the P_{CT} (preferentially conserved targeting) value, which shows the conservative degree of target binding sites (23). Context score (22) is also output in the list. Context score shows the predicted efficacy of targeting. When a gene/transcript has more than one miRNA binding site, the summation of P_{CT} is used for ranking. The percentile of a binding site is the percent point of the binding site in a descending order of context score of all binding sites of a miRNA.

Stem-loop reverse transcription TaqMan polymerase chain reaction

Novel miRNAs predicted by miRDeep* from small RNAseq were validated using the miRNA-specific stem-loop reverse transcription (SL-RT) TaqMan polymerase chain reaction (PCR) (33). Briefly, RT was carried out on 1 µg of DNAse I (Life Technologies)-treated total RNA and using 1 unit of superscript III as per the manufacturer's instructions (Life Technologies). cDNA synthesis was performed at 42°C. Oligonucleotides used in RT for each miRNA are detailed in Supplementary Table S1. TaqMan PCR was carried out using the TaqMan Universal PCR Master Mix (Applied Biosystems, Mulgrave, VIC, Australia) as per the manufacturer's instructions, but with modifications as described previously (34), and using primer annealing temperatures of 55°C. TaqMan probes and PCR primers are detailed in Supplementary Table S1. Gene expression was quantified against a standard curve of synthetic mature miRNA oligonucleotides (Sigma-Aldrich, Castle Hill, NSW, Australia). Random hexamer generated cDNA was used as a negative control for SL-RT TaqMan PCR specificity and for assessing *RPL32* gene expression for normalizing miRNA expression. Quantitative RT–PCR for *RPL32* was carried out using the Sybr Green PCR mastermix (Applied Biosystems) as described previously (35). Data are represented as the standard error mean from three independent RNA extractions.

Qualitative RT–PCR

Qualitative RT–PCR was carried out to detect pre-miRNA as further validation of miRDeep*-redicted mature miRNA. RT was performed as per SL-RT TaqMan PCR, but using random hexamer oligonucleotides, and cDNA synthesis was performed at 50°C. PCR was carried out using platinum taq DNA polymerase (Life Technologies), a 60°*C primer annealing temperature and 40 cycles of PCR. A negative RT control was used to ensure that genomic DNA was not amplified as pre-miRNAs are intron less. All primer sequences are detailed in Supplementary Table S2. Data are represented as the mean ± standard error from RNA extracted from three independently treated cells. Two tailed t-testing was used to assess for significant differences in expression.

RESULTS

Benchmarking of miRDeep* sensitivity and specificity in detecting mature miRNA against miRDeep, miRanalyzer, miRTRAP and MIReNA

Androgens are important mediators of prostate cancer physiology. Thus, we have used the most widely studied androgen-responsive prostate cancer model, the LNCaP cell line, to investigate miRNA that are downstream of the AR signaling axis. In total, 3 937 673 and 7 791 811 small RNAseq reads were obtained from mock (ethanol)- and R1881 (synthetic androgen)-treated LNCaP cells, respectively. The miRNA predictive capacity of miRDeep* was then compared against the popular miRDeep2, miRDeep, miRanalyzer, miRTRAP and MIReNA. miRNA prediction algorithms. The miRNA detection threshold for miRDeep2, miRDeep and miRDeep* was set at the previously recommended score of 0 (16). In total, miRDeep*-predicted 208 (mock dataset) and 237 (R1881) miRNAs, whereas miRDeep2-predicted 272 (mock dataset) and 320 (R1881 dataset) miRNAs, miRDeep-predicted 203 (mock dataset) and 235 (R1881 dataset) miRNAs and miRanalyzer-predicted 1063 (mock dataset) and 1321 (R1881 dataset) miRNA (Figure 3A, Table 1 and Supplementary Table S3). For miRDeep*, miRDeep2 and miRDeep, the majority of the highest scoring miRNAs, and thus more likely to be bona fide miRNA, were known miRNA that are already in the miRBase database v13. In contrast, miRanalyzer had a higher proportion of novel miRNA with high number of reads when compared with miRDeep*, miRDeep2 and miRDeep. MiRTRAP predicted similar number of miRNAs with miRDeep2, but the precision is lower than miRDeep. MIReNA predicted a small number of miRNAs, but a few of them are not in miRBase (36). Both the precision and recall of MIReNA are lower than miRDeep*.

An analysis was then performed to compare the capability of miRDeep* in detecting known miRNA using the same datasets with that of the miRDeep, miRDeep2, miRanalyzer, miRTRAP and MIReNA programs. MiRanalyzer predicted approximately five times more miRNA from our LNCaP dataset compared with

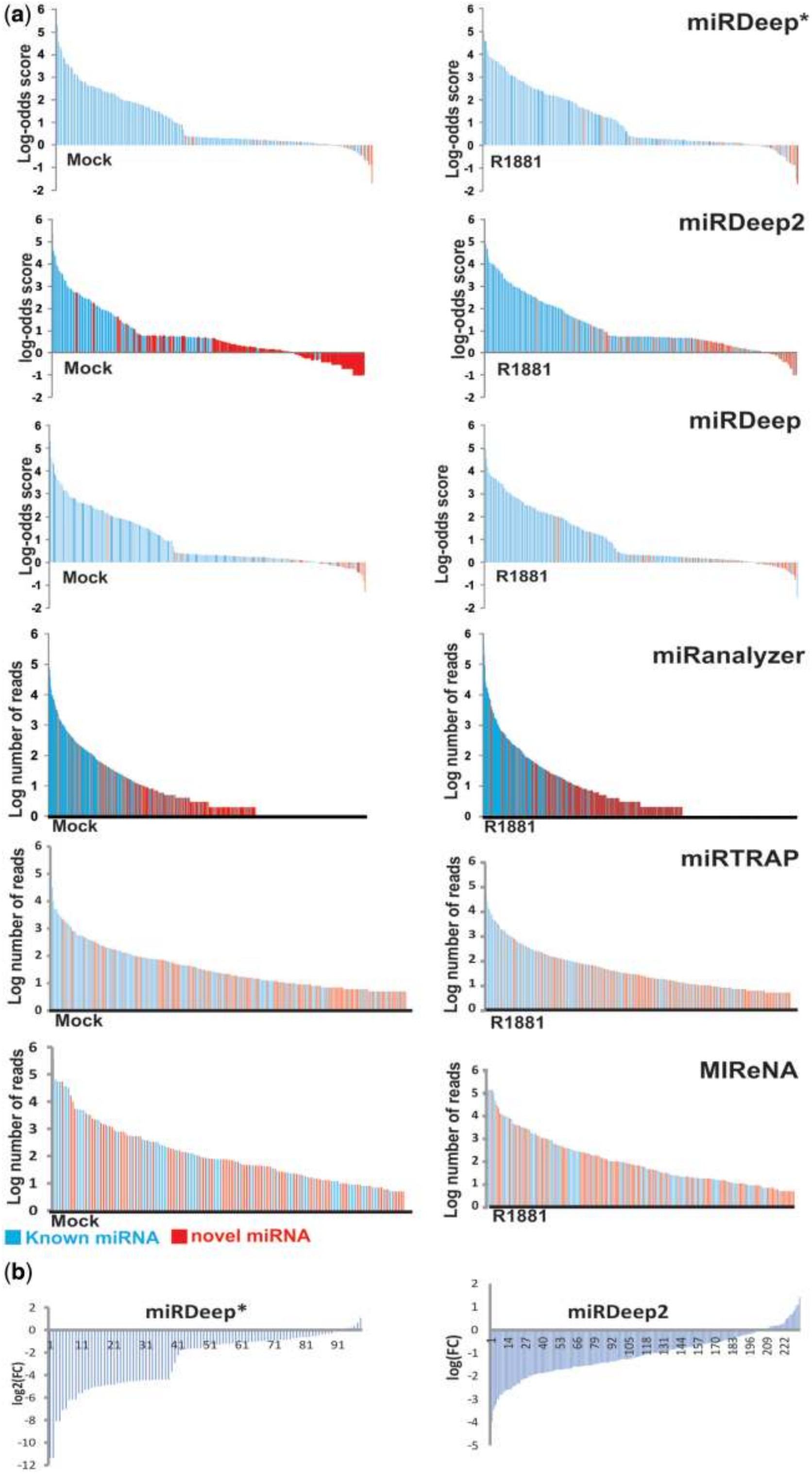


Figure 3. (**a**) Vertical axis represents the log scores of predicted miRNAs obtained from miRNA prediction algorithm. The horizontal axis lists predicted miRNAs in descending order of score. Green lines show the predicted miRNAs that are known from miRBase. The red lines show the predicted miRNAs not listed in miRBase (i.e. novel miRNAs). The vertical axis is log scale, because the score difference across predicted miRNAs is huge. Since miRanalyzer, miRTRAP and MIReNA do not provide scores, the vertical axes represent the number of reads in the predicted miRNA. The number of predicted miRNAs shown in the figure is reduced to match the other algorithms. (**b**) Vertical axis represents the log fold changes after anti-dicer induced. The horizontal axis lists predicted miRNAs in descending order of folder changes.

Table 1. Comparative analysis of the sensitivity and specificity of miRDeep*

	miRDeep*	miRDeep2	miRDeep	miRanalyzer	miRTRAP	MIReNA
Mock treated						
Number of known miRNA found in raw RNAseq reads[a,b]	240	240	240	240	240	240
Number of predicted miR	208	272	203	1063	298	162
Number of predicted miR in miRBase[b]	173	208	164	215	158	77
Number of predicted novel miR	35	64	39	848	140	85
Precision	83.17%	76.47%	80.79%	20.23%	53.02%	47.53%
Recall	72.08%	86.67%	68.33%	89.58%	65.83%	32.08%
R1881 treated						
Number of known miRNA found in raw RNAseq reads[a,b]	285	285	285	285	285	285
Number of predicted miR	237	320	235	1321	345	190
Number of predicted miR in miRBase[b]	192	229	180	261	175	88
Number of predicted novel miR	45	91	55	1060	170	102
Precision	81.01%	71.56%	76.60%	19.76%	50.72%	46.32%
Recall	67.37%	80.35%	63.16%	91.58%	61.40%	30.88%

Precision = Number of predicted miRNA in miRBase/Number of predicted miRNA. Recall = Number of predicted miRNA in miRBase/RNAseq reads found in miRBase.

[a] Where at least 5 raw RNAseq reads map to each known miRNA in miRBase version 13.

[b] miRBase version 13.

miRDeep*, miRDeep2 and miRDeep (Table 1). However, the precision, which is measured relative to known miRNA (37), of miRanalyzer-predicted miRNA was only 19.76 and 20.23%. This indicates that ~80% of the miRNA predicted by miRanalyzer are novel and without any validation. miRDeep* and miRDeep predicted much fewer miRNA; however, the precision for both algorithms was >70%, with miRDeep* having a 2.9 and 5.8% higher precision for the mock- and R1881-treated datasets, respectively (Table 1). miRDeep* also slightly outperformed miRDeep in detecting validated miRNA using the same datasets with 5.5 and 6.7% higher recall. Interestingly, miRDeep2 had lower precision (76.47 and 71.56%) compared with the original miRDeep and miRDeep*. Conversely, miRDeep2 has higher recall ratio (86.67 and 80.35%) than miRDeep and miRDeep* in terms of recall (Table 1).

Like most miRNA prediction tools, we have set a threshold to stop mapping a read when the number of mapped genomic loci exceeds this threshold (default threshold is 101). As shown in Table 2, uniquely mapped reads of all predicted miRNAs from our LNCaP dataset comprised more than three quarters of the predicted RNAseq reads. More than 99.8% reads are mapped into one, two or three genomic loci.

We tested miRDeep* using the miRDeep2 small RNAseq dataset (17) to further assess miRdeep* in predicting novel miRNA. Significantly, these data were generated before and after inducing anti-dicer in MCF-7 cells to turn off the miRNA biogenesis pathway. The miRNA predicted by miRDeep* has a lower average log(FC) compared with miRDeep2, which demonstrates that these novel miRNAs are more likely to be generated from the miRNA biogenesis pathway (Figure 3B and Table 3). This is further supported as the percentage of miRNA with a negative fold change after Dicer knockdown, which indicates that mature miRNA are no longer produced, is greater than that of miRDeep2. However, miRDeep2 was able to predict more novel miRNA than miRDeep*.

We have also assessed whether miRDeep* is able to detect novel miRNA that was previously reported in two prostate cancers from a small RNAseq dataset. Of the 28 novel miRNAs detected by Martens-Uzunova *et al.* (11), miRDeep* was able to detect seven of these in our LNCaP prostate cancer cells (Supplementary Table S4).

We have further evaluated the performance of miRDeep* on 7 of the 406 small RNAseq datasets derived from a breast cancer study (38). The seven datasets analyzed composed of four different stages of the disease. The precision of all but one of these seven datasets is >83% (Supplementary Table S7). Similarly, the recall of all seven datasets is >80% (Supplementary Table S7). Most of the miRNAs that were undetected by miRDeep* have very low number of reads, and collectively, these results are consistent with our LNCaP analyses.

Validation of miRDeep*-predicted miRNA

Five novel miRNAs predicted by miRDeep* (2 miRNA with the highest score, novel 1 and 2 and three of the higher expressing miRNA with more than 100 reads, novels 3, 4 and 5) from our LNCaP RNAseq dataset were chosen for validation using SL-RT TaqMan PCR. However, one of these novel miRNAs (novel 5) could not be amplified, including the synthesized positive control oligonucleotide. Consequently, further validation was only carried out for the other four novel miRNA predicted by miRDeep*. The highly expressed hsa-miR-29a, an androgen-regulated miRNA in prostate cancer cells, and hsa-let-7a-2 expression were also assessed by SL-RT TaqMan PCR. The novel 1 miRNA maps to Chromosome 17 and is located within and antisense of the *AATK* and *hsa-miR-338* genes (Figure 4a). The novel 2 miRNA maps to Chromosome X and is located within the *EDA* gene (Figure 4a). The novel 3 miRNA is located antisense of *NT5DC1*, and the novel 4 miRNA is located within and antisense of *ARHGAP44* (Figure 4a). The novel 1 and 2 miRNAs were predicted by all methods (miRDeep*, miRDeep2, miRDeep, miRanalyzer,

Table 2. Multiple mapped reads

Mapping	MOCK		R1881	
	Reads	Percentage	Reads	Percentage
1	582 762	82.14	864 909	74.22
2	44 545	6.28	79 788	6.85
3	81 175	11.44	219 510	18.84
4	675	0.10	919	0.08
5	25	0.00	34	0.00
6	174	0.02	39	0.00
7	50	0.01	0	0.00
8	58	0.01	22	0.00
9	14	0.00	15	0.00
10	0	0.00	9	0.00
11	0	0.00	38	0.00
12	0	0.00	36	0.00
13	0	0.00	5	0.00
14	0	0.00	0	0.00
15	0	0.00	0	0.00
16	0	0.00	24	0.00
17	15	0.00	0	0.00
18	0	0.00	39	0.00

Table 3. Difference between before and after anti-dicer for novel miRs identified by miRDeep* and miRDeep2

	Average log(FC)	miRs with negative FC (%)	Number of novel miRs
miRDeep*	−2.74	85.70	98
miRDeep2	−1.01	74.79	234

miRTRAP and MIReNA) in both mock and R1881 datasets (Supplementary Table S5). The novel 3 miRNA was predicted by miRDeep* and miRDeep, and the novel 4 miRNA was predicted by only miRDeep* in both mock and R1881 datasets and by miRDeep in mock dataset.

SL-RT TaqMan PCR was able to specifically detect the mature miRNA as TaqMan PCR on random hexamer-generated cDNA using the same quantity of synthetic miRNA resulted in a much higher C_t value compared with PCR on cDNA generated from SL oligonucleotides targeting each novel miRNA (Supplementary Table S6). Of the six miRNAs assessed by SL-RT TaqMan PCR, only hsa-miR29a was significantly upregulated by R1881 in LNCaP cells, which is consistent with previous studies (Figure 4a). Conversely, the novel 2 miRNA was significantly down-regulated by R1881 (Figure 4a). The pre-miRNA secondary structure of these novel miRNAs and their expression levels (for ethanol-treated cells) are shown in Figure 4b. The expression of pre-miRNA for each of these novel miRNA using qualitative RT–PCR (Figure 4c) demonstrates that they are synthesized and processed in the classical pri- and pre-miRNA biogenesis pathways.

DISCUSSION

miRNAs are a class of ncRNA that are increasingly found to be important in many aspects of normal developmental and disease processes. Indeed, the mechanisms through which miRNA mediate the post-transcriptional effects appear to extend beyond the classical translational repression and RNA degradation pathways. Recent studies have discovered that miRNA might also be involved in regulating the actions of ribonucleoprotein complexes in a mechanism that is independent of the canonical RISC pathway (1). The levels of a number of miRNA have been reported to be deregulated in many complex diseases, such as cancer, and at least 26 altered miRNA profiles have already been described in prostate cancer alone (4). In fact, a recent study has identified a panel of miRNA, which can be used to differentiate between normal and prostate cancer cells (39). These recent discoveries highlight the importance of identifying miRNA that are expressed in normal and disease contexts so that therapeutic and disease management strategies can be employed.

We have taken advantage of the recent advances made in next-generation RNAseq to profile the androgen-regulated small-RNA transcriptome in prostate cancer cells. We investigated the miRNA expression profile in our small RNAseq dataset on the basis of the importance of miRNA in prostate cancer, and the role of androgens in prostate cancer progression (40). However, to use the existing miRNA prediction tools—miRDeep and its varieties (16, 41, 42)—we have to pre-process the data using several other application tools such as a read alignment tool and a RNA secondary structure tool. Moreover, some miRNA, such as miR-25 and miR-200a, were found to be highly expressed in our LNCaP dataset by miRDeep* but were not detected by miRDeep due to improper excision of the pre-miRNA region in those algorithms. Importantly, Martens *et al.*, discovered novel miRNA in their study, and miRDeep* was able to detect seven of these novel miRNA in our LNCaP small RNAseq data.

miRDeep* is a standalone, one-click, integrated, RNAseq-based miRNA prediction tool. This application does not require any pre-installed computational tools, such as a RNA secondary structure tool or a genome alignment tool. These tools are built into miRDeep* using pure Java coding. Therefore, this application can be easily executed on any platform (Linux, Windows and MacOS). Although there are several RNAseq-based miRNA computational tools, we provide the first truly intuitive and most user-friendly tool that has a graphics display output showing the RNAseq reads relative to the pre-miRNA secondary structure. This display also illustrates miRNA isoforms, including the copy numbers for each variant. This application also accepts aligned reads in SAM/BAM format as input files. One of the major advantages of miRDeep* in reducing analysis time is that, unlike other web-based algorithms, miRDeep* is a standalone application where all the data are readily accessible to the user, making it easy to adjust the source code and to change the miRNA prediction strategy. Moreover, miRDeep* is not affected by other users because this standalone application runs on a local machine. Sequence alignment and RNA secondary structure modeling of 3 million reads takes only 3 h to perform

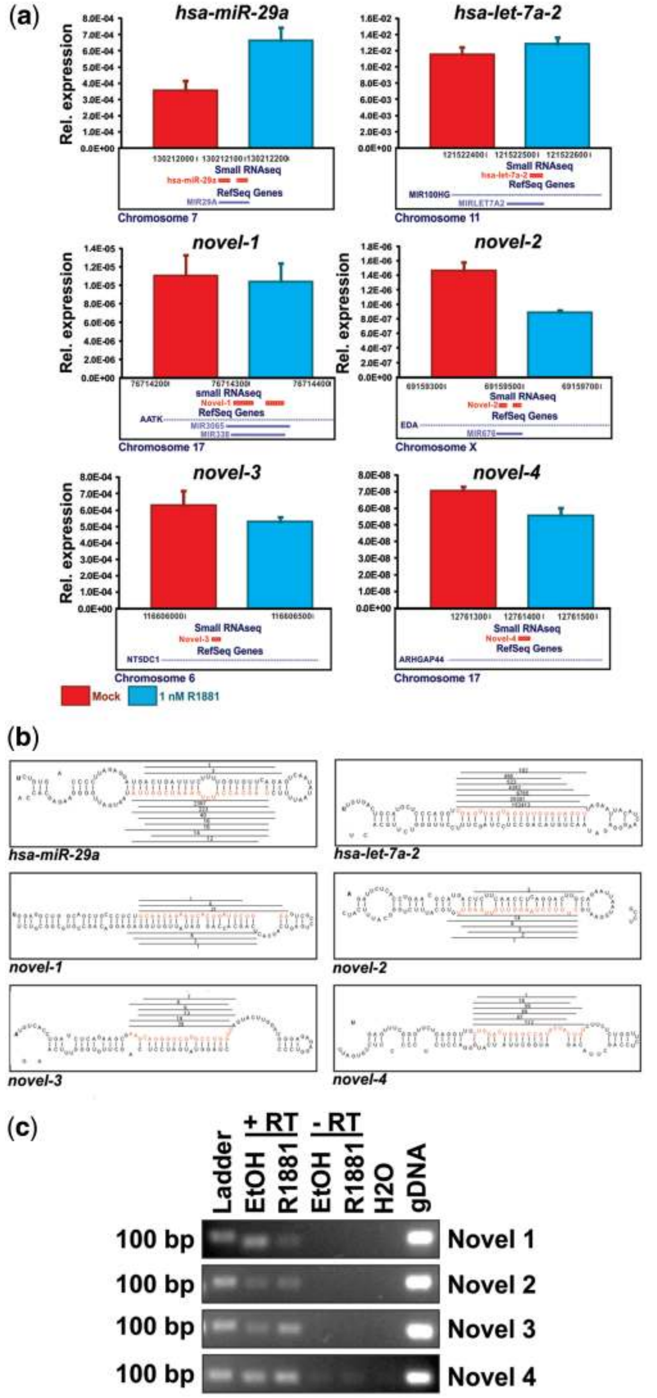


Figure 4. Validation of miRDeep*-predicted novel miRNA. (**a**) Validation was carried out using stem-loop TaqMan PCR for four novel miRNA (novel 1–4), and hsa-miR-29a and hsa-let-7a-2, which served as positive controls. Small RNAseq was performed on RNA that was extracted from LNCaP prostate cancer cells that were treated with ethanol (mock) or androgen (1 nM R1881) for 48 h. Represented is the mean expression ± SE from three independently treated cells. (**b**) Representation of the pre-miRNA secondary structure predicted by miRDeep*, including the location of the mature miRNA (red nucleotides). Also shown are the locations and number of RNAseq reads (lines above and below the predicted pre-miRNA). (**c**) Qualitative RT–PCR showing expression of novel 1, 2, 3, and 4 pre-miRNA. No reverse transcriptase (−RT) controls were included to ensure that genomic DNA (gDNA) was not amplified. gDNA and a negative template (H_2O) were included as positive and negative controls, respectively.

on miRDeep* using a PC windows XP with 2.5 GHz CPU and 4 G RAM, whereas a similar analysis can take up to 12 h using the web-based miRanalyzer and miRTRAP and MIReNA.

Many of the current small RNAseq-based miRNA prediction tools have been influenced by the highly effective miRDeep tool (16, 43). For example, deepBase that has a repository of 2038 predicted miRNAs (44) used 185 publically available small RNAseq experiments and the miRDeep algorithm to compile the database. seqBuster (45) provides a web-based and standalone application to analyse known miRNA variants. miRanalyzer (46) is a web-based application that finds known and novel miRNAs from sequence data and uses a random forest machine learning algorithm to predict novel miRNAs. miRNAKey (47) and DSAP (48) determine the differential known miRNA expression from deep sequence data. More recently, the miRDeep2 program was introduced, which improved on the original miRDeep algorithm as well as other functions such as graphical output. Coincidentally, we were independently developing miRDeep* in parallel with miRDeep2, using similar algorithms to identify the precursor miRNA, and extra functionalities such as a graphical output. However, miRDeep* has slightly different approaches including a dynamic graphical output as opposed to miRDeep2's static output.

We have benchmarked miRDeep* against miRDeep, miRDeep2, miRTRAP, mIReNA and miRanalyzer using the same small RNAseq reads. Our analyses was carried out under no *a priori* knowledge that the dataset was derived from human prostate cancer cells and so attempted to find all possible miRNAs. miRanalyzer was able to detect approximately 5-fold more miRNA from the same dataset compared with miRDeep*, miRDeep2 and miRDeep; however, the precision for miRanalzer was only ~20%. miRDeep* had slightly better precision (less than 10%) and recall (less than 5%) compared with miRDeep. Interestingly, miRDeep2 had lower precision compared with the original miRDeep for our dataset. However, the recall of miRDeep2 was much better than miRDeep* (mock: ~27%; R1881: ~18%) and miRDeep (mock: ~30%; R1881: ~23%). The improved predictive ability of miRDeep* was further supported by using a dataset whereby the classical miRNA biogenesis pathway was turned off through Dicer knock-down. Using this dataset, bona fide mature miRNA are less likely to be predicted in cells where Dicer was knocked out. Our results demonstrated that miRDeep* is less likely to result in false-positive predictions of miRNA compared with miRDeep2. MiRTRAP detected similar numbers of miRNA from the same dataset with miRDeep2, but its precision was lower than miRDeep2. However, miRTRAP has a better performance than MIReNA in terms of precision and recall in our datasets. Finally, we have analysed seven other small RNAseq datasets from breast cancers using miRDeep* and found that the performance (precision and recall) is similar to the results from our LNCaP datasets.

Of the four miRNAs chosen for validation by SL-RT TaqMan PCR, one (novel 4) was predicted solely by miRDeep*. Thus, miRDeep* is both more precise at predicting miRNA from small RNAseq reads and having greater capacity to detect novel miRNA compared with miRDeep and miRDeep2. Moreover, miRDeep* is able to predict miRNA that are formed from the classical pri- and pre-miRNA biogenesis pathways as shown by the detection of pre-miRNA in our validation experiments.

In conclusion, we have developed miRDeep* that is a useful tool for predicting novel miRNA from small RNAseq. miRDeep* is the first application for predicting miRNAs without any pre-installed software and has better predictive capabilities compared with miRDeep and miRDeep2. miRDeep* also offers numerous user-friendly and functional advantages, such as full Java coding, which makes it suitable for use with many operating systems, as well as displaying RNAseq reads and the number of reads relative to the predicted pre-miRNA. Importantly, this tool provides the target prediction for both known and novel miRNAs, unlike most tools, which do not provide target prediction for novel miRNAs. Given the role of miRNA in many biological contexts, including prostate cancer, we propose that miRDeep* will be an extremely useful tool for researchers.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1–7.

FUNDING

Queensland Smart Futures Premier's Fellowship (to C.C.N.); Prostate Cancer Foundation of Australia (PCFA PG25), Australian Government Department of Health and Ageing. Funding for open access charge: Queensland University of Technology.

Conflict of interest statement. None declared.

REFERENCES

1. Yang,J.S. and Lai,E.C. (2011) Alternative miRNA biogenesis pathways and the interpretation of core miRNA pathway mutants. *Mol. Cell.*, **43**, 892–903.
2. Cloonan,N., Wani,S., Xu,Q., Gu,J., Lea,K., Heater,S., Barbacioru,C., Steptoe,A.L., Martin,H.C., Nourbakhsh,E. *et al.* (2011) MicroRNAs and their isomiRs function cooperatively to target common biological pathways. *Genome Biol.*, **12**, R126.
3. Djuranovic,S., Nahvi,A. and Green,R. (2011) A parsimonious model for gene regulation by miRNAs. *Science*, **331**, 550–553.
4. Catto,J.W., Alcaraz,A., Bjartell,A.S., De Vere White,R., Evans,C.P., Fussel,S., Hamdy,F.C., Kallioniemi,O., Mengual,L., Schlomm,T. *et al.* (2011) MicroRNA in prostate, bladder, and kidney cancer: a systematic review. *Eur. Urol.*, **59**, 671–681.
5. Wang,D. and Tindall,D.J. (2011) Androgen action during prostate carcinogenesis. *Methods Mol. Biol.*, **776**, 25–44.
6. Perner,S., Mosquera,J.M., Demichelis,F., Hofer,M.D., Paris,P.L., Simko,J., Collins,C., Bismar,T.A., Chinnaiyan,A.M., De Marzo,A.M. *et al.* (2007) TMPRSS2-ERG fusion prostate cancer: an early molecular event associated with invasion. *Am. J. Surg. Pathol.*, **31**, 882–888.
7. Tomlins,S.A., Laxman,B., Varambally,S., Cao,X., Yu,J., Helgeson,B.E., Cao,Q., Prensner,J.R., Rubin,M.A., Shah,R.B. *et al.* (2008) Role of the TMPRSS2-ERG gene fusion in prostate cancer. *Neoplasia*, **10**, 177–188.
8. Wang,W.L., Chatterjee,N., Chittur,S.V., Welsh,J. and Tenniswood,M.P. (2011) Effects of 1alpha,25 dihydroxyvitamin D3 and testosterone on miRNA and mRNA expression in LNCaP cells. *Mol. Cancer*, **10**, 58.
9. Ribas,J., Ni,X., Haffner,M., Wentzel,E.A., Salmasi,A.H., Chowdhury,W.H., Kudrolli,T.A., Yegnasubramanian,S., Luo,J., Rodriguez,R. *et al.* (2009) miR-21: an androgen receptor-regulated microRNA that promotes hormone-dependent and hormone-independent prostate cancer growth. *Cancer Res.*, **69**, 7165–7169.
10. Watahiki,A., Wang,Y., Morris,J., Dennis,K., O'Dwyer,H.M., Gleave,M. and Gout,P.W. (2011) MicroRNAs associated with metastatic prostate cancer. *PLoS One*, **6**, e24950.
11. Martens-Uzunova,E.S., Jalava,S.E., Dits,N.F., van Leenders,G.J., Moller,S., Trapman,J., Bangma,C.H., Litman,T., Visakorpi,T. and Jenster,G. (2012) Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene*, **31**, 978–991.
12. Ostling,P., Leivonen,S.K., Aakula,A., Kohonen,P., Makela,R., Hagman,Z., Edsjo,A., Kangaspeska,S., Edgren,H., Nicorici,D. *et al.* (2011) Systematic analysis of microRNAs targeting the androgen receptor in prostate cancer cells. *Cancer Res.*, **71**, 1956–1967.
13. Hackenberg,M., Rodriguez-Ezpeleta,N. and Aransay,A.M. (2011) miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments. *Nucleic Acids Res.*, **39**, W132–W138.
14. Hendrix,D., Levine,M. and Shi,W. (2010) miRTRAP, a computational method for the systematic identification of miRNAs from high throughput sequencing data. *Genome Biol.*, **11**, R39.
15. Mathelier,A. and Carbone,A. (2010) MIReNA: finding microRNAs with high accuracy and no learning at genome scale and from deep sequencing data. *Bioinformatics*, **26**, 2226–2234.
16. Friedlander,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.
17. Friedlander,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
18. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
19. Li,R., Yu,C., Li,Y., Lam,T.W., Yiu,S.M., Kristiansen,K. and Wang,J. (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.
20. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
21. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.
22. Grimson,A., Farh,K.K., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,D.P. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
23. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
24. Morin,R.D., O'Connor,M.D., Griffith,M., Kuchenbauer,F., Delaney,A., Prabhu,A.L., Zhao,Y., McDonald,H., Zeng,T., Hirst,M. *et al.* (2008) Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, **18**, 610–621.
25. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
26. Moxon,S., Schwach,F., Dalmay,T., Maclean,D., Studholme,D.J. and Moulton,V. (2008) A toolkit for analysing large-scale plant small RNA datasets. *Bioinformatics*, **24**, 2252–2253.
27. Friedlander,M.R., Chen,W., Adamidi,C., Maaskola,J., Einspanier,R., Knespel,S. and Rajewsky,N. (2008) Discovering

microRNAs from deep sequencing data using miRDeep. *Nat. Biotechnol.*, **26**, 407–415.

28. Hofacker,I.L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res.*, **31**, 3429–3431.
29. Maragkakis,M., Vergoulis,T., Alexiou,P., Reczko,M., Plomaritou,K., Gousis,M., Kourtis,K., Koziris,N., Dalamagas,T. and Hatzigeorgiou,A.G. (2011) DIANA-microT Web server upgrade supports Fly and Worm miRNA target prediction and bibliographic miRNA to disease association. *Nucleic Acids Res.*, **39**, W145–W148.
30. Enright,A.J., John,B., Gaul,U., Tuschl,T., Sander,C. and Marks,D.S. (2003) MicroRNA targets in *Drosophila*. *Genome Biol.*, **5**, R1.
31. Bandyopadhyay,S. and Mitra,R. (2009) TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples. *Bioinformatics*, **25**, 2625–2631.
32. Wang,X. (2006) Systematic identification of microRNA functions by combining target prediction and expression profiling. *Nucleic Acids Res.*, **34**, 1646–1652.
33. Chen,C., Ridzon,D.A., Broomer,A.J., Zhou,Z., Lee,D.H., Nguyen,J.T., Barbisin,M., Xu,N.L., Mahuvakar,V.R., Andersen,M.R. *et al.* (2005) Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res.*, **33**, e179.
34. Cheng,A., Li,M., Liang,Y., Wang,Y., Wong,L., Chen,C., Vlassov,A.V. and Magdaleno,S. (2009) Stem-loop RT-PCR quantification of siRNAs in vitro and in vivo. *Oligonucleotides*, **19**, 203–208.
35. Kaushal,A., Myers,S.A., Dong,Y., Lai,J., Tan,O.L., Bui,L.T., Hunt,M.L., Digby,M.R., Samaratunga,H., Gardiner,R.A. *et al.* (2008) A novel transcript from the KLKP1 gene is androgen regulated, down-regulated during prostate cancer progression and encodes the first non-serine protease identified from the human kallikrein gene locus. *Prostate*, **68**, 381–399.
36. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36**, D154–D158.
37. Tempel,S. and Tahi,F. (2012) A fast ab-initio method for predicting miRNA precursors in genomes. *Nucleic Acids Res.*, **40**, e80.
38. Farazi,T.A., Horlings,H.M., Ten Hoeve,J.J., Mihailovic,A., Halfwerk,H., Morozov,P., Brown,M., Hafner,M., Reyal,F., van Kouwenhove,M. *et al.* (2011) MicroRNA sequence and expression analysis in breast tumors by deep sequencing. *Cancer Res.*, **71**, 4443–4453.
39. Martens-Uzunova,E.S., Jalava,S.E., Dits,N.F., van Leenders,G.J., Moller,S., Trapman,J., Bangma,C.H., Litman,T., Visakorpi,T. and Jenster,G. (2011) Diagnostic and prognostic signatures from the small non-coding RNA transcriptome in prostate cancer. *Oncogene*, **31**, 978–991.
40. Massard,C. and Fizazi,K. (2011) Targeting continued androgen receptor signaling in prostate cancer. *Clin. Cancer Res.*, **17**, 3876–3883.
41. Yang,J.H., Zhang,X.C., Huang,Z.P., Zhou,H., Huang,M.B., Zhang,S., Chen,Y.Q. and Qu,L.H. (2006) snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res.*, **34**, 5112–5123.
42. Zhu,E., Zhao,F., Xu,G., Hou,H., Zhou,L., Li,X., Sun,Z. and Wu,J. (2010) mirTools: microRNA profiling and discovery based on high-throughput sequencing. *Nucleic Acids Res.*, **38**, W392–W397.
43. Friedlander,M.R., Mackowiak,S.D., Li,N., Chen,W. and Rajewsky,N. (2012) miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res.*, **40**, 37–52.
44. Yang,J.H., Shao,P., Zhou,H., Chen,Y.Q. and Qu,L.H. (2010) deepBase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res.*, **38**, D123–D130.
45. Pantano,L., Estivill,X. and Marti,E. (2010) SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells. *Nucleic Acids Res.*, **38**, e34.
46. Hackenberg,M., Sturm,M., Langenberger,D., Falcon-Perez,J.M. and Aransay,A.M. (2009) miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res.*, **37**, W68–W76.
47. Ronen,R., Gan,I., Modai,S., Sukacheov,A., Dror,G., Halperin,E. and Shomron,N. (2010) miRNAkey: a software for microRNA deep sequencing analysis. *Bioinformatics*, **26**, 2615–2616.
48. Huang,P.J., Liu,Y.C., Lee,C.C., Lin,W.C., Gan,R.R., Lyu,P.C. and Tang,P. (2010) DSAP: deep-sequencing small RNA analysis pipeline. *Nucleic Acids Res.*, **38**, W385–W391.