4298–4305 *Nucleic Acids Research, 2012, Vol. 40, No. 10* *Published online 28 January 2012*
doi:10.1093/nar/gks043

Performance comparison and evaluation of software tools for microRNA deep-sequencing data analysis

Yue Li[1], Zhuo Zhang[2], Feng Liu[1], Wanwipa Vongsangnak[1], Qing Jing[2,3,]* and Bairong Shen[1,4,]*

[1]Center for Systems Biology, Soochow University, Suzhou 215006, China, [2]Department of Cardiology, Changhai Hospital, Shanghai 200433, China, [3]Key Laboratory of Stem Cell Biology, Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences & Shanghai Jiao-Tong University School of Medicine, Shanghai, China and [4]Department of Bioinformatics, Medical College, Soochow University, Suzhou 215123, China

Received October 11, 2011; Revised December 22, 2011; Accepted January 10, 2012

ABSTRACT

With the development of next-generation sequencing (NGS) techniques, many software tools have emerged for the discovery of novel microRNAs (miRNAs) and for analyzing the miRNAs expression profiles. An overall evaluation of these diverse software tools is lacking. In this study, we evaluated eight software tools based on their common feature and key algorithms. Three deep-sequencing data sets were collected from different species and used to assess the computational time, sensitivity and accuracy of detecting known miRNAs as well as their capacity for predicting novel miRNAs. Our results provide useful information for researchers to facilitate their selection of the optimal software tools for miRNA analysis depending on their specific requirements, i.e. novel miRNAs discovery or miRNA expression profile analysis of sequencing data sets.

INTRODUCTION

MicroRNAs (miRNAs) are a class of small RNAs with a small number of nucleotides, i.e. 18–25 bp, which have essential roles in a variety of cellular processes, such as organism development, metabolism, immunological responses and tumorigenesis (1,2). miRNAs can either repress mRNA translation or induce the cleavage of mRNA targets via hybridization with the 3′-untranslated region of target mRNAs (1,3,4). Small RNA cloning methods were used to identify novel miRNAs (5), but sequencing technology has evolved rapidly and next-generation sequencing (NGS) appears to be very promising for miRNAs detection, because it provides the major advantages of high-throughput sequencing (6,7) with very high speed and reduced cost. Several studies have successfully used NGS for the discovery of novel miRNAs, especially for those that are difficult to detect (7,8) at a low abundance.

Since the application of NGS to miRNA detection, many sequencing software tools have been developed to support miRNA data analysis. These include miRDeep (9), miRanalyzer (10), miRExpress (6), miRTRAP (11), DSAP (12), mirTools (7), MIReNA (8), miRNAkey (13) and mireap which can be accessed at http://sourceforge.net/projects/mireap/.

miRDeep and mireap were early software tools used for analyzing deep-sequencing small RNA data sets generated by NGS. However, they are limited to organisms where known reference genomes are available. miRanalyzer can be applied widely in different organisms via a web server tool that can handle 11 different organisms. miRExpress can be used when no genome sequencing is available (6). miRTRAP can be used in case of an existed gene annotation file format is gff (11). DSAP is an automated multitask web service that facilitates comparative miRNA analysis, such as differential expression, cross-species distribution and phylogenetic distribution (12). mirTools provides detailed annotation for each known miRNA (7) and it allows the determination of the relative expression level of all miRNAs, which can be illustrated using a scatter plot where red dots represent differentially expressed miRNAs (7). Finally, MIReNA can predict miRNAs and pre-miRNAs in the following data sets: known miRNA sequencing; deep-sequencing data; putative pre-miRNAs, possibly including miRNA candidates; and long sequencing, including potential miRNAs (8). However, the current study is only limited to deep-sequencing data analysis. miRNAkey has a user-friendly graphical user interface (GUI) that can be used for visualizing differentially expressed miRNAs in paired samples (13). The common features of the nine software tools are summarized in Supplementary Table S1A.

*To whom correspondence should be addressed. Tel:/Fax: +86 512 65110951; Email: bairong.shen@suda.edu.cn
Correspondence may also be addressed to Qing Jing. Tel: +86 21 54920609; Fax: +86 21 54920612; Email: qjing@sibs.ac.cn

Given this brief description of the individual software tools available, it would be useful to consider each program's capacity in terms of computational time, sensitivity, and accuracy as well as its relevance for predicting novel miRNAs. Thus, we aimed to compare different miRNA sequencing software tools to further evaluate their capabilities. The eight sequencing software tools were tested using public deep-sequencing data sets derived from three different genomes, i.e. human (*Homo sapiens*), chicken (*Gallus gallus*) and worm (*Caenorhabditis elegans*). This study provides useful information for researchers when selecting the optimal software tools for miRNA analysis depending on their specific requirements and it provides a reference for computational biologists developing novel software tools.

miRNA sequencing software tools must address two important issues when identifying miRNAs, i.e. mapping deep-sequencing reads of genomes and predicting the secondary structures of each mapped locus. Detailed features of miRNAs prediction using each software tool are summarized in Supplementary Table S1B.

To address the first issue, alignment algorithms used in sequence mapping were considered as fundamental components of these software tools. We classified each program based on the alignment algorithm and the year the software tool was released, as shown in Figure 1A. MIReNA, miRDeep and miRTRAP employed the MegaBLAST/BLAST algorithm, so they were classified together. MIReNA uses the approach introduced in miRDeep (8,9) based on mapping deep-sequencing reads of the genome using MegaBLAST, while miRTRAP (11) uses the BLAST approach. SOAP2 is employed in mirTools, which can reduce computer memory usage and increase the alignment speed (14). The SEQ-EM algorithm is applied by miRNAkey to optimize the distribution of multiple aligned reads among the observed miRNAs, before mapping using a reference database of known miRNAs with the Burrows–Wheeler Aligner (BWA) (13). Both DSAP and miRExpress use the word-match and Smith–Waterman algorithm, so they are placed in the same class. This algorithm is more appropriate for handling a large number of sequences on web servers (6,12,15,16). miRanalyzer has two alignment options based on whether reads have adapter sequences (10). There is no documentation available for mireap, so it is grouped as an unclassified algorithm.

The algorithm used for predicting the secondary structures of mapped loci was also compared. In Figure 1B, each software tool is classified with respect to the algorithm used for generating the optimal secondary structure and the year of the software's release. miRTRAP, miRDeep and mirTools use RNAfold and Bayes' theorem to evaluate the secondary structure of candidate RNAs. Up to 100 and 150 nt of the genomic sequence flanking the mapped locus in individual reads is extracted and folded using RNAfold by miRTRAP and mirTools, respectively (7,11). Bayes' theorem is used by miRDeep for scoring potential miRNA precursors (9). miRExpress retrieves cross-species sequence information from the UCSC Genome Browser (17) to determine the conservation of putative miRNAs (6). To detect new miRNAs, miRanalyzer applies a machine-learning approach based on a WEKA implementation of the random forest learning scheme where the number of trees is set to 100 (10). MIReNA searches for miRNA sequences by exploring a multidimensional space defined using only five parameters to characterize acceptable miRNA precursor (8). The interface features of each software tool are indicated in Figure 1, including the web server, local server and GUI.

MATERIALS AND METHODS

Short reads and data sets

Caenorhabditis elegans deep-sequencing data from 454 sequencing technology were obtained from the NCBI GEO database, which was produced by combining five sequencing reactions from five different mixed-stage samples (accession no. GSE5990) (8). *Gallus gallus* deep-sequencing data was generated from small RNA libraries prepared with Day 5 (CE5) chicken embryos (NCBI GEO database accession no. GSE10636) (11). miRNA sequencing data from undifferentiated human

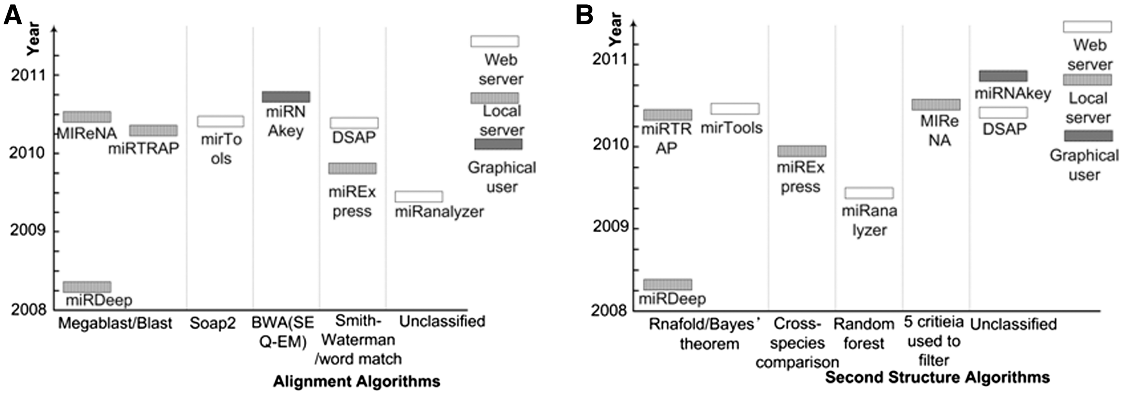


Figure 1. Summary of miRNA sequencing software tools. Software tools were clustered according to their prediction strategies. Different symbols are used to distinguish tools suited to different types of running platforms. (**A**) The algorithms used by miRNA sequencing programs for alignment. (**B**) The algorithm used by miRNA sequencing programs for secondary structure prediction.

embryonic stem cells were downloaded from ftp://ftp03.bcgsc.ca/public/hESC_miRNA/; H9_day0_trimmed_and_mapped_with_counts.txt.gz (16). Known miRNA sequences and their genome locations were downloaded from miRBase version 16 (http://www.mirbase.org/).

Program implementation

All miRNA sequencing software tools were run with the default or recommended settings from a server equipped with four 2.4 GHz Intel(R) Xeon(R) 4 CPUs, with four cores in each CPU and 32 GB of RAM. The operating system was Ubuntu 8.04.4 using version of X_86 64 bits.

Prediction system assessment

To evaluate the performance of the software tools, the following quantities were calculated: the number of miRNAs correctly predicted (true positives, TP), the number of pseudo-miRNAs incorrectly predicted as real miRNAs (false positives, FP) and the number of miRNAs incorrectly predicted as pseudo-miRNAs (false negatives, FN). We used the following measures to evaluate the performance of the software tools.

Sensitivity (Sen) = TP/(TP + FN)

Accuracy (Acc) = TP/(TP+FP+FN).

DNA sequences and gene annotation

Caenorhabditis elegans, *G. gallus* and *H. sapiens* sequences and their gene annotations were retrieved from UCSC (http://hgdownload.cse.ucsc.edu/downloads.html). March 2004 (WS120/ce2), May 2006 (WUGSC 2.1/galGal3) and February 2009 (GRCh37/hg19) assemblies were used, respectively.

Secondary structure prediction

mfold was used for secondary structure prediction (http://mfold.rna.albany.edu/?q = mfold/RNA-Folding-Form).

RESULTS

Computational time

The computational time required by each software tool was determined with data sets for three different organisms (*H. sapiens*, *G. gallus* and *C. elegans*) based on the algorithms of the miRNA sequencing software tools. The computational time results are shown in Figure 2. Many web tools use different computational resources that are hidden in the internet, so we only focused on local servers. Compared with software tools used on the local server, mireap took less computational time (10 min for *G. gallus* and 43 min for *H. sapiens*) compared with miRDeep (10 days for *C. elegans* and one month for *H. sapiens*) and MIReNA (10 days for *C. elegans* and more than one month for *H. sapiens*).

Sensitivity

To evaluate the sensitivity if miRNA identification using different miRNA sequencing software tools, we tested their performance when applied to three public deep-sequencing data sets of *H. sapiens, G. gallus* and *C. elegans*. The miRNAs of each species found in miRBase were defined as the positive set while others comprised the negative set. This criterion allowed us to compare the sensitivity of different software tools in a consistent way. Potential genuine miRNAs may not be registered in the miRBase, so a more precise characterization of a positive set based on miRBase was needed. It was reasonable to group the candidates as true miRNAs if they were predicted as miRNAs using three or more software tools. miRNA candidates in the human, chicken and worm are shown in Supplementary Tables S2A–C, respectively. Thus, a combination of miRBase with these miRNA candidates was viewed as an alternative definition of the positive set (extended positive set).

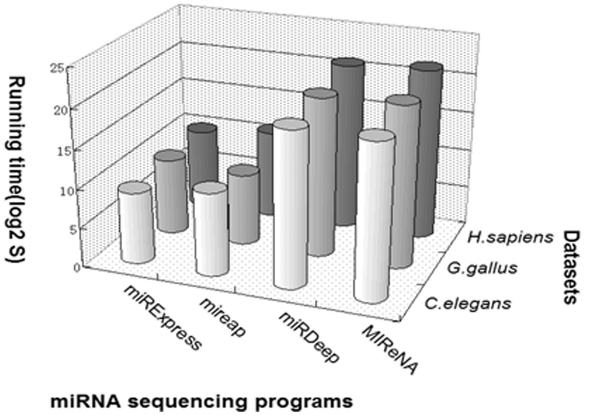


Figure 2. Computational time. The time cost for each miRNA sequencing software tool with different data sets.

The number of predicted miRNAs was compared with those in miRBase and the percentage of known miRNAs identified by each software tool is shown in Figure 3. With respect to each genome, miRExpress and DSAP had the highest success (72.1 and 71.2%, respectively) when predicting miRNAs from *C. elegans*. miRExpress, DSAP and mirTools covered 77–80% of miRNAs for *G. gallus*. miRanalyzer had the highest success of 60.6% when predicting miRNAs for *H. sapiens*.

miRExpress produced the highest success when predicting miRNAs for *G. gallus* and *C. elegans*, while it also had relatively high success with *H. sapiens*. DSAP also had relatively higher success when predicting miRNAs in all three organisms. mirTools was ranked third when compared with the other software tools. mirTools is a user-friendly web server tool with a short computational time of 5–6 h when completing each data set run. miRanalyzer had the highest prediction success with *H. sapiens* and a satisfactory high success with *C. elegans*. miRanalyzer is also a web server tool, but it requires 1–2 weeks to complete each data set run. Compared with mirTools, miRanalyzer was obviously more time consuming.

The number of predicted miRNAs was compared with the extended positive set, as shown in Figure 4.

miRExpress had the highest success (71.19 and 78.5%, respectively) when predicting miRNAs for *C. elegans* and *G. gallus*. mireap had the highest success of 59.85% for *H. sapiens*. These results suggest that different software tools were suited to predicting miRNAs in specific data sets.

Accuracy

Accuracy is an important issue when predicting miRNAs. Obviously, predicted miRNAs with less false positives are preferred.

miRBase was used as a reference standard. The number of predicted miRNAs was compared with the total number of predicted miRNAs and the percentage of known miRNAs identified by each software tool is shown in Figure 5. miRDeep had the highest success of 97.41% when predicting *C. elegans*. mirTools had the highest success of 90.69% when predicting miRNAs for *G. gallus*. miRExpress had the highest success of 87.65% when predicting *H. sapiens*. mirTools had the highest success of 90.69% when predicting *G. gallus*, 95.2% when predicting *C. elegans* and 84.1% when predicting *H. sapiens*. The next best were miRDeep, MIReNA, miRExpress and mireap.

The extended positive set was then used and the success rate is shown in Figure 6. miRExpress had the highest success of 97.11% when predicting *C. elegans*. mirTools had the highest success of 90.81% with *G. gallus,* while miRanalyzer had the highest success of 100% in *H. sapiens*. Thus, the performance accuracy when predicting miRNAs also depended on the data set used.

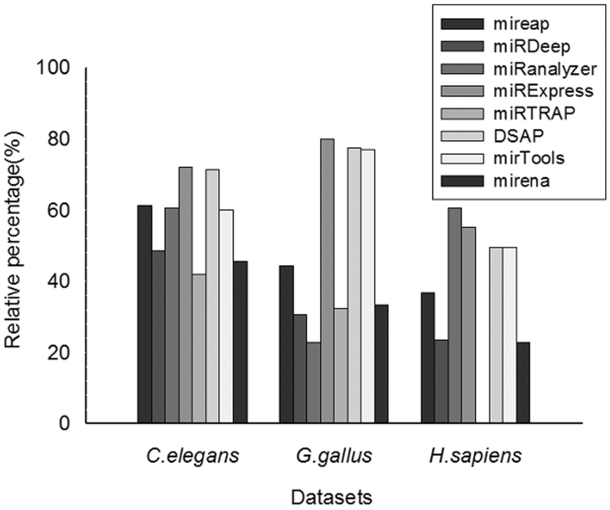


Figure 3. Comparison of the sensitivity of various software tools when predicting known miRNAs. Programs reported different numbers of miRNAs when run with their default or recommended settings using the same data sets. The percentage of predicted miRNAs in miRBase using different data sets is shown.

Finally, a comparison of the grouped software tools was made. Eight software tools were separated into two groups according to their running platform, i.e. web server or local server. Figure 7A shows the success when detecting known miRNAs with miRDeep, mireap, MIReNA and miRTRAP, which are based on local servers. mireap had the highest mean success in detecting known miRNAs from the three data sets (89.4, 83.5 and 82.5%, respectively).

Figure 7B shows the success with miRanalyzer, miRExpress, DSAP and mirTools, which are based on web servers. miRExpress had the highest success of 92.3% with *C. elegans* and 93.8% with *G. gallus*. miRanalyzer had the highest success of 81.8% with *H. sapiens*.

Venn diagram

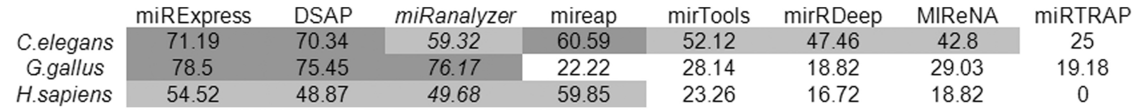
The Venn diagrams of miRDeep, mireap and MIReNA when predicting known miRNAs are shown in Figure 8A–C for the three organisms, respectively. The highest overlap was at the intersection of predicting *C. elegans*, whereas it was comparatively lower with *G. gallus* and *H. sapiens*. The predicting of known miRNAs was more clustered with *C. elegans,* whereas they were more discrete with *H. sapiens* when using these three programs.

Predicting novel miRNAs

It was difficult to make an unbiased evaluation of the capability of the different software tools when predicting novel miRNAs, because it was not possible to identify all the genuine miRNAs. Thus, we adopted a practicable approach where small RNA sequences were categorized as true miRNAs is they were repeatedly identified as miRNAs using distinct groups of programs.

Fourteen novel *G. gallus* miRNAs predicted using three or more software tools are shown in Supplementary Table S2A. mireap and miRDeep had the highest frequency. Selected secondary structures of potential novel miRNAs are shown in Figure 9.

Fifteen novel *H. sapiens* miRNAs predicted by the three software tools are shown in Supplementary Table S2B. MIReNA had the highest frequency. miRDeep and mirTools were ranked second and third. Gene annotations are shown in Supplementary Table S2B. According to UCSC Genome Browser, 13.3, 26.7, and 60% of reads were in exon, intronic and intergenic regions, respectively.



	miRExpress	DSAP	*miRanalyzer*	mireap	mirTools	mirRDeep	MIReNA	miRTRAP
C.elegans	71.19	70.34	*59.32*	60.59	52.12	47.46	42.8	25
G.gallus	78.5	75.45	*76.17*	22.22	28.14	18.82	29.03	19.18
H.sapiens	54.52	48.87	*49.68*	59.85	23.26	16.72	18.82	0

Figure 4. Comparison of the sensitivity of various software tools when predicting known miRNAs. Programs reported different numbers of miRNAs, when run with their default or recommended settings using the same data sets. The percentage success when predicting miRNAs in miRBase and the miRNAs identified by three or more software tools from different data sets are shown. Entries are shaded with black and white gradients, where black represents the highest percentage and white the lowest.

Selected secondary structures of potential novel miRNAs are shown in Supplementary Figures S1–S3.

Three novel *C. elegans* miRNAs predicted using three or four software tools are shown in Supplementary Table S2C. MIReNA had the highest frequency, followed by mireap, miRDeep and miRanalyzer. Eight novel miRNAs were predicted by MIReNA, three of which were in miRBase version 17 but not in version 16, while another four had hairpin structures predicted by mfold. Secondary structures of potential novel miRNAs are shown in Supplementary Figure S4. Gene annotations are shown in Supplementary Table S2C. According to the Wormbase Gene Annotation in the UCSC Genome Browser, three reads were in intronic or intergenic regions.

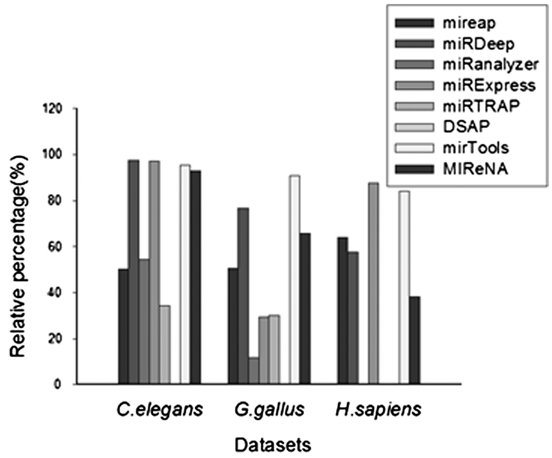


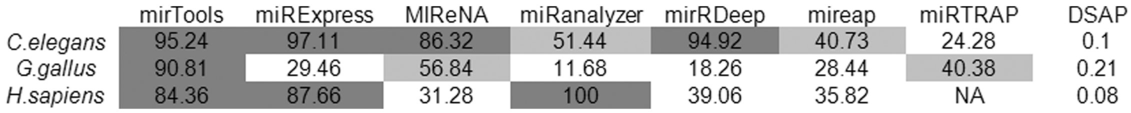
Figure 5. Comparison of the accuracy of various software tools when predicting known miRNAs. Programs reported different numbers of miRNAs when run with their default or recommended settings using the same data sets. The percentage of predicted miRNAs in miRBase is compared with the total number of predicted miRNAs with different data sets.

DISCUSSION

Performance evaluation

To identify true miRNAs, eight software tools excluded other RNA fragments by rigorously comparing a read with known rRNAs, scRNAs, snRNAs, snoRNAs, tRNAs or mRNAs. miRDeep, MIReNA, miRanalyzer and DSAP aligned the reads to Rfam, RepBase or mRNA sequencing (8–10,12). With miRTRAP, mirTools or mireap, reads are separated as valid miRs based on a measurement or classification (7,11). If there is a match, the read is excluded as a miRNA. miRExpress can accept reads by aligning sequences with known miRNAs (6).

miRExpress gave the best sensitivity performance when detecting known miRNAs, which may be due to miRExpress constructing miRNA expression profiles by aligning sequences with known miRNAs (6). DSAP ranked the second when detecting known miRNAs, but it predicted 169705 miRNAs of *C. elegans*, 201 733 miRNAs of *G. gallus* and 737 516 miRNAs of *H. sapiens*. Therefore, it is likely to produce a large number of false positives. miRanalyzer and mirTools



	mirTools	miRExpress	MIReNA	miRanalyzer	mirRDeep	mireap	miRTRAP	DSAP
C.elegans	95.24	97.11	86.32	51.44	94.92	40.73	24.28	0.1
G.gallus	90.81	29.46	56.84	11.68	18.26	28.44	40.38	0.21
H.sapiens	84.36	87.66	31.28	100	39.06	35.82	NA	0.08

Figure 6. Comparison of the accuracy of various software tools when predicting known miRNAs. Programs reported different numbers of miRNAs when run with their default or recommended settings using the same data sets. The percentage success when predicting miRNAs in miRBase and the miRNAs identified by three or more software tools compared with the total number of predicted miRNAs are shown for different data sets. Entries are shaded with black and white gradients, where black represents the highest percentage and white the lowest.

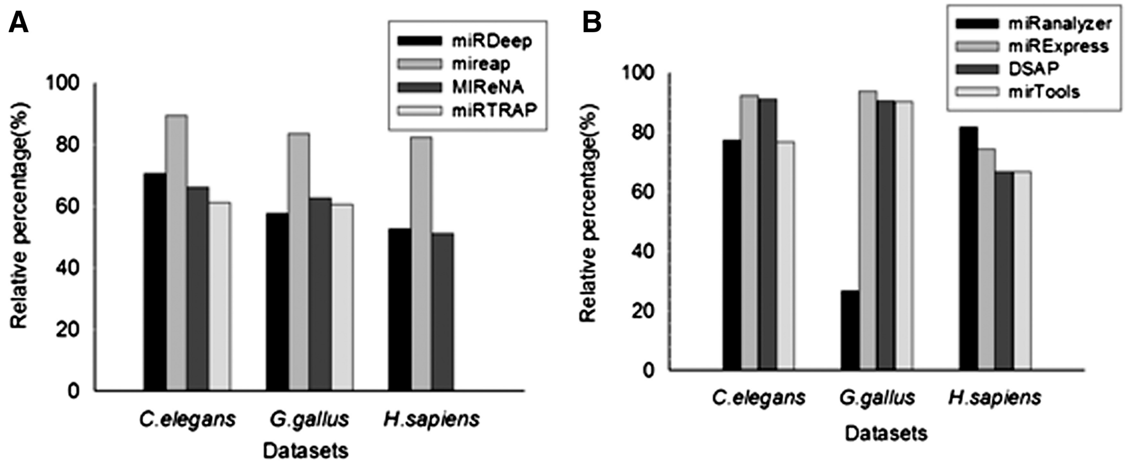


Figure 7. Comparison of the detection rate of grouped software tools when predicting known miRNAs. (**A**) Local server programs: miRDeep, mireap, MIReNA and miRTRAP; (**B**) web server programs: miRanalyzer, miRExpress, miRExpress and mirTools.

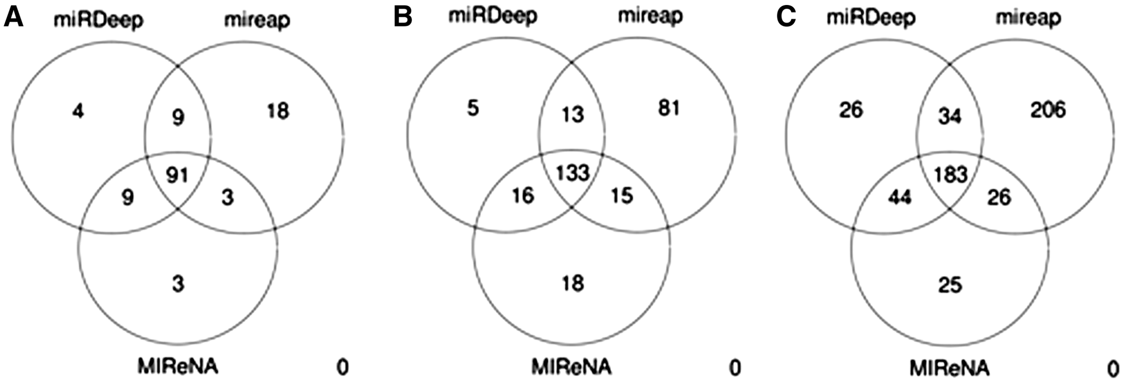


Figure 8. Venn diagram of miRDeep, mireap and MIReNA when predicting known miRNAs. (**A**) *Caenorhabditis elegans*, (**B**) *G. gallus* and (**C**) *H. sapiens*.

had the best sensitivity performance with *C. elegans*. mirTools had the best performance with *G. gallus* while miRanalyzer had the best performance with *H. sapiens*. Compared with mirTools and miRanalyzer, mirTools was faster because SOAP2 was used to make alignments, which reduces computer memory usage and increases the alignment speed at an unprecedented rate (7,14).

miRTRAP predicted 98 known miRNAs and 142 novel miRNAs for *C. elegans*, 176 known miRNAs and 409 novel miRNAs for *G. gallus*, but zero miRNAs for *H. sapiens*. This software was applied to the simple chordate *Ciona intestinalis* and it identified nearly 400 putative miR loci (11)

mfold and RNALogo are applied in the discovery of novel miRNAs by miRExpress, but few novel miRNAs were predicted by miRExpress. Therefore, mirTools had the best accuracy performance when predicting miRNAa in all three species and it was also the best software tool for detecting known miRNAs in terms of sensitivity and accuracy.

When the eight software tools are separated into two groups based on their running platform, mireap had the highest success when recovering known miRNAs from the three data sets on a local server and it also ran faster.

The Venn diagram shows that the predicted known miRNAs were more overlapped in *C. elegans* whereas they differed more in *H. sapiens* when using these three programs. In general, highly expressed miRNAs tend to be functionally important and evolutionarily conserved, whereas the low expression human miRNA genes that comprise ~30% of currently annotated genes are almost free of selective pressure (18). Most miRNAs in this group may only occasionally enter the small RNA biosynthesis pathway. miRNA candidates identified only by individual programs are not likely to stably demonstrate well-characterized features of miRNAs, so they elude repeated observation with multiple program searches.

It is difficult to assess software tools based on their capacity to predict novel miRNAs, but command line tools gave better performance than web server tools. The best recommended combinations of software tools for a particular data set are shown in Table 1. MIReNA is the first choice for nematode and mammal data sets. Combinations of mireap, miRDeep and miRanalyzer can be used with nematode. miRDeep and mirTools can also be used with mammals. In vertebrates, mireap was the first choice, while miRDeep and mirTools can also be integrated. miRDeep has better performance when predicting novel miRNAs for *C. elegans*, because it used *C. elegans* data for parameter estimation (9). MIReNA had better performance when predicting novel miRNAs with *C. elegans* and *H. sapiens* compared with miRDeep, because it searches for miRNA sequences by exploring a multidimensional space defined using only five (physical and combinatorial) parameters characterizing acceptable pre-miRNAs. It detects new miRNAs based on homology with known miRNAs or deep-sequencing data (8). A major feature of MIReNA is the capacity to adapt the search to specific species, possibly characterized by the specific properties of their miRNAs and pre-miRNAs (8).

Selecting the ‘right’ software tool for different tasks

Various software tools have emerged for miRNA deep-sequencing data analysis. Thus, it is important to select the ‘best’ tool. Before selecting ‘suitable’ software tools, the organisms gene annotation format must be available. miRExpress can be used with any organism, but mfold and RNALogo must be used in the discovery of novel miRNAs. miRanalyzer, DSAP, miRExpress, mirTools and miRNAkey can be used in a comparative mode. DSAP can perform comparative miRNAomics for mature miRNAs or miRNA families with up to a maximum of five jobs. Appropriate software tool should be selected based on the input and output requirements. However, we recommend the software tools in Table 1 for different data types, according to the performance.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2 and Supplementary Figures 1–4.

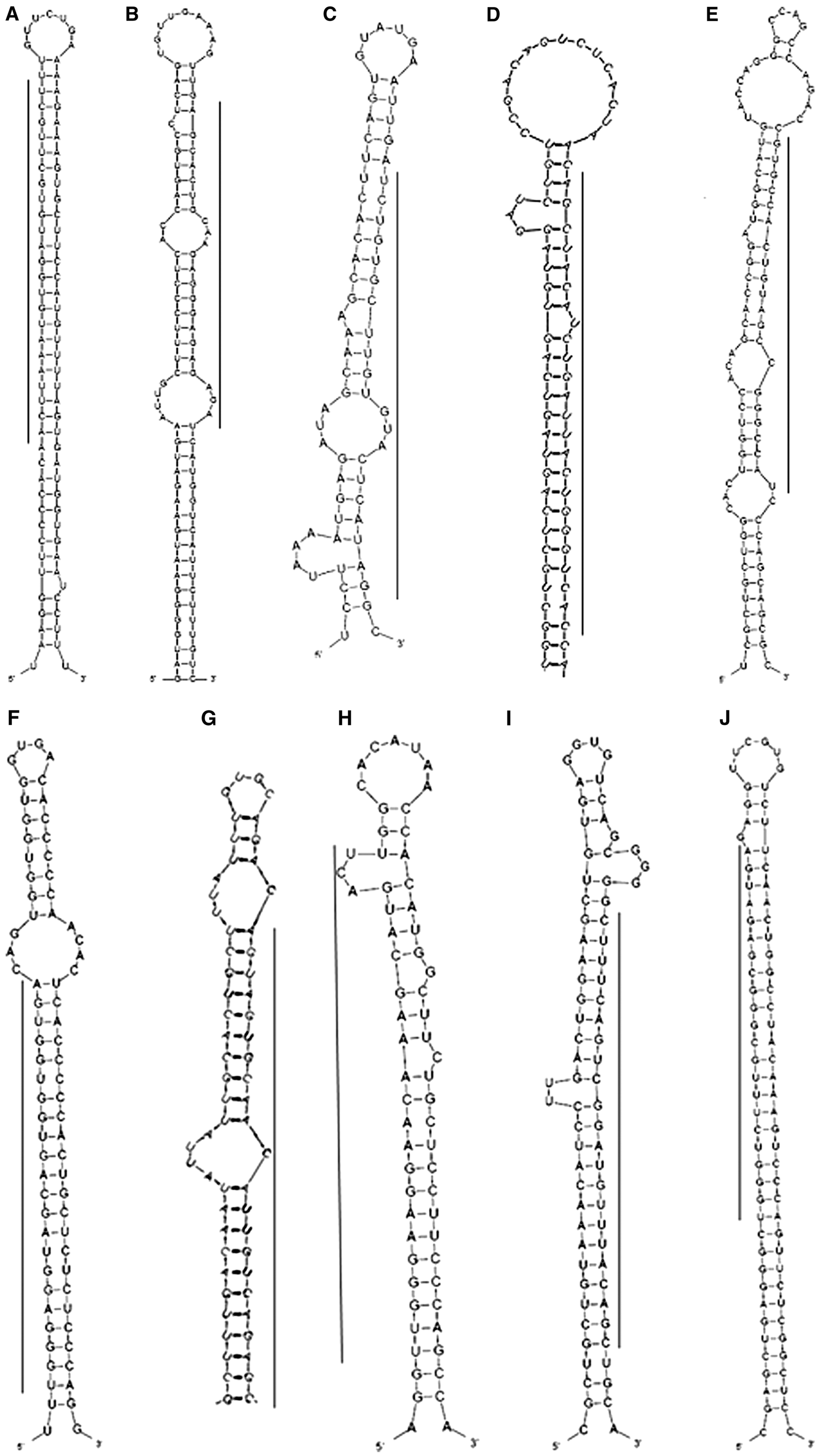


Figure 9. Secondary structures of novel *G. gallus* miRNAs predicted using mfold. (**A**) reads seq_11973_x6; (**B**) reads seq_37970_x22; (**C**) reads seq_10960_x6; (**D**) reads seq_24186_x3; (**E**) reads seq_127592_x1; (**F**) reads seq_14978_x4; (**G**) reads seq_11344_x6; (**H**) reads seq_5255_x14; (**I**) reads seq_1568_x67; (**J**) reads seq_65731_x1.

Table 1. Recommended software in predicting novel miRNAs for each data set

Clade	Genome used in this article	Recommended software in predicting novel miRNAs		
		1	2	3
Nematode	*Caenorhabditis elegans*	MIReNA	mireap miRDeep	miRanalysis
Vertebrate	*Gallus gallus*	mireap	miRDeep	miRTRAP
Mammal	*Homo sapiens*	MIReNA	miRDeep	mirTools

ACKNOWLEDGEMENTS

We thank Biao Chen, Chao Xu, Wen-Yu Zhang, Qing-Hui Li, Yi-Fei Tang, Jing Shang and Jing Wang for helpful discussions.

FUNDING

Funding for open access charge: The National Natural Science Foundation of China (31170795, 91029703); International S&T Cooperation Program of Suzhou (SH201120); Major State Basic Research Development Program of China (2007CB947002, 2010CB945600).

Conflict of interest statement. None declared.

REFERENCES

1. He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.
2. Lu,J., Getz,G., Miska,E.A., Alvarez-Saavedra,E., Lamb,J., Peck,D., Sweet-Cordero,A., Ebert,B.L., Mak,R.H., Ferrando,A.A. *et al*. (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.
3. Brennecke,J., Stark,A., Russell,R.B. and Cohen,S.M. (2005) Principles of microRNA-target recognition. *PLoS Biol.*, **3**, e85.
4. Kim,V.N. and Nam,J.W. (2006) Genomics of microRNA. *Trends Genet.*, **22**, 165–173.
5. Lagos-Quintana,M., Rauhut,R., Lendeckel,W. and Tuschl,T. (2001) Identification of novel genes coding for small expressed RNAs. *Science*, **294**, 853–858.
6. Wang,W.C., Lin,F.M., Chang,W.C., Lin,K.Y., Huang,H.D. and Lin,N.S. (2009) miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression. *BMC Bioinformatics*, **10**, 328.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. Farrar,M. (2007) Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics*, **23**, 156–161.
16. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
17. Kuhn,R.M., Karolchik,D., Zweig,A.S., Wang,T., Smith,K.E., Rosenbloom,K.R., Rhead,B., Raney,B.J., Pohl,A., Pheasant,M. *et al*. (2009) The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res.*, **37**, D755–D761.
18. Liang,H. and Li,W.H. (2009) Lowly expressed human microRNA genes evolve rapidly. *Mol. Biol. Evol.*, **26**, 1195–1198.