BMC Bioinformatics

Methodology article

Open Access

miRAS: a data processing system for miRNA expression profiling study

Feng Tian[1,2,3], Huayue Zhang[2,3], Xinyu Zhang[2,3], Chi Song[1,3], Yongjing Xia[2], Yiqing Wu[2] and Xiangjun Liu*[1,2,3]

Address: [1]Department of Biological Science and Biotechnology, Tsinghua University, Beijing 100084, People's Republic of China, [2]Institutes of Biomedicine, Tsinghua University, Beijing 100084, People's Republic of China and [3]Ministry of Education Key Laboratory of Bioinformatics, Tsinghua University, Beijing 100084, People's Republic of China

Email: Feng Tian - tianfeng@mails.tsinghua.edu.cn; Huayue Zhang - zhanghuayue06@mails.tsinghua.edu.cn; Xinyu Zhang - zhangxinyu04@mails.tsinghua.edu.cn; Chi Song - songchi@mails.tsinghua.edu.cn; Yongjing Xia - yjxia@mail.tsinghua.edu.cn; Yiqing Wu - yqwu@mail.tsinghua.edu.cn; Xiangjun Liu* - frankliu@tsinghua.edu.cn

* Corresponding author

Published: 4 August 2007

BMC Bioinformatics 2007, **8**:285 doi:10.1186/1471-2105-8-285

Received: 30 April 2007
Accepted: 4 August 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/285

Abstract

Background: The study of microRNAs (miRNAs) is attracting great considerations. Recent studies revealed that miRNAs play as important regulators of gene expression and some even as cancer players or inhibitors. Many studies try to discover new miRNAs and reveal the miRNA expression profile in cancer using a SAGE-based total RNA clone method. However, the data processing of this method is labor-intensive with several different biological databases and more than ten data processing steps involved.

Results: With miRAS, miRNAs and possible miRNA candidates contained in the submitted sequencing data were obtained together with their expression profile. The functions of known and predicted miRNAs were then analyzed by miRNA target prediction followed by target gene annotations. Finally, to extract the biological significance of the miRNAs in the samples, further annotations of the known miRNA and target genes were performed by collecting the public expression datasets of miRNA and target genes in normal and cancer tissues.

Conclusion: We introduce a web-based analysis platform called miRNA Analysis System (miRAS), for processing and analyzing of the sequence data obtained from the total RNA clone method. The system was built on generalizing the study of a liver cancer cell line total RNA sequencing project. miRAS is freely available on the web.

Background

MicroRNAs (miRNAs) are a class of small non-coding RNAs that regulate gene expression by binding to their target mRNAs and triggering either protein translation repression or RNA degradation [1]. Recent studies show that some miRNAs are located at fragile sites and genomic regions involved in cancers [2]. The aberrant expression of miRNA genes could lead to human disease, including cancer [3-6], and are regarded as potential biomarkers for cancer diagnosis [7,8]. The roles miRNAs play have been demonstrated in a few cancer types including breast cancer [9], lung cancer [10] and chronic lymphocytic leukemia [2,11,12], while the roles of miRNA in other cancers remain largely unknown.

There are several approaches of studying miRNAs and their expression profiles, including Northern blotting and real-time PCR assay. There are also available high-throughput methods such as oligonucleotide miRNA microarray analysis [13-15], bead-based flow-cytometric technique [8], and SAGE-based miRAGE [16]. miRNA microarray analysis is a commonly used high-throughput technique for the assessment of previously discovered miRNAs. With the SAGE-based technique, such as miRAGE, the expression profiles of known miRNAs could be retrieved together with the unknown ones which are possible miRNA candidates.

For gene expression SAGE studies [17,18], there exist several well developed methods for data analysis together with web services provided, such as SAGEmap [19] and SAGE Genie [20]. For miRNA-related SAGE, however, the data analysis is much more complicated. The extracted tags have to be compared with various RNA databases in addition to mRNA sequences. The tags also need to be mapped to the human genome and to be analyzed for precursors with thermodynamically stable hairpin structures. This is a very troublesome process and current users have to refer to several different databases to retrieve related biologically significant data [16]. To aid the processing and data analysis of this method, we constructed a web-based system, named miRNA Analysis System (miRAS). The expression profile of known miRNAs in submitted sequences were returned and compared with public dataset using Fisher's exact test. Public available datasets of known miRNAs expression in liver were collected for the annotation of miRNA expression in liver. Several public available gene expression datasets were included to reveal differentially expressed genes in liver cancer and normal liver tissues. The differentially expressed miRNAs and genes are highlighted and the relationship between miRNAs and genes is shown according to miRNA target prediction.

Results and Discussion

Users could upload the raw sequencing data and specify the sequencing parameters through the web interface. The known miRNAs and possible miRNA candidates will be analyzed together with their expression profiles. The target genes predicted by miRNA target prediction software are provided together with the annotation information. To demonstrate the biological significance of the retrieved miRNAs, the profiles of public datasets of known miRNAs and target genes were collected and included in the annotation.

The miRAS system provides an easy and friendly way for scientists to analyze and process raw miRNA sequence data to obtain new miRNA candidates. It also provides tools for the annotations of the predicted miRNAs.

Conclusion

In this work, we established a web-based analysis platform for miRNAs, called miRAS [21], to analyze the miRNA expression in specific tissue and to predict and study the possible miRNA candidates. The differentially expressed miRNAs that target differentially expressed genes are retrieved together with miRNA and target gene annotation, to uncover the biological significance. Currently it supports liver cancer genes, while in the future, the analysis platform is planned to be expanded to support other cancers and to integrate all public available expression data of the miRNAs and genes in cancer and normal tissues.

Methods

The work flow of miRAS is diagrammed in Figure 1. Details of the major steps of the system are described in the following sections.

Preliminary processing of sequence data

Multiple sequence data of RNA sequencing in multi-fasta format could be uploaded for analysis directly. The raw sequencing data in trace file format could be compressed in tar, tar.gz, zip, or rar format, and uploaded.

If the trace data of total RNA sequencing result is selected, the sequence and quality of the sequence data will be checked by PHRED [22,23], which reads DNA sequence trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files. Sequences with low quality base call values, to be removed, are detected by PHRED with default parameters, which explore modified Mott trimming algorithm, an error probability cutoff value of 0.05 and a minimum segment length of 20 bases. The produced high-quality reads are then compared to the vector sequence consensus using the cross-match program, the mostly used software of finding vector regions on sequencing reads. Parameters of -minmatch 12 -minscore 20 are used as default, while users could also apply their own parameters through the web interface. A standard vector database provided by PHRAP package is used as default and it can also be specified by the users. The detected vector sequences contained in the sequence reads will be removed. The adapter sequences contained in the miRNA clone sequences are checked and removed similarly, using the cross-match program with parameters (-minmatch 8 -minscore 14) for combined adaptors and (-minmatch 8 -minscore 8) for single adapters.

Extraction of known miRNAs and primary new candidate miRNAs

After removal of the adapter sequences, redundant RNA segments are removed and the copy numbers of each segment are recorded. The resulting short RNA segments are searched against a known human miRNA dataset (miRNA Registry Release 9.1) [24,25] to retrieve known miRNAs.

The expressions of the known miRNAs in given tissues are compared with public expression data to obtain differentially expressed or tissue specifically expressed miRNAs. Other non-coding RNA (ncRNA) are identified in a similar way, by searching against a ncRNA database, which was built by combining several public RNA databases, including rRNA database from Ribosomal Database Project (RDP) [26], ncRNA database from NONCODE [27] and Regulatory noncoding RNAs database [28]. 1 nt miss match is allowed in retrieving the blast output for the tolerance of mistake in DNA sequencing. Those not included in the known miRNAs or other ncRNA databases are treated as primary miRNA candidates.

Pre-miRNA retrieval and secondary miRNA candidates screening

The primary miRNA candidates identified from the previous step are aligned to the human genome. For each aligned segment, the 99 nucleotides upstream to the alignment position on the genome, the sequence of the primary candidate miRNA itself, and the 99 nucleotides downstream to it on the genome are concatenated and used as a possible precursor sequence of the primary candidate. The secondary structures of the precursor sequences are predicted with the mfold program [29] and RNAfold [30], and checked with two criteria of miRNA precursors: 1) whether the candidate precursor sequence has a characteristic hairpin secondary structure and 2) whether a single miRNA molecule accumulates one arm of a hairpin precursor molecule [31]. Support vector machine (SVM) [32] is applied to classify real vs. pseudo pre-miRNAs.

miRNA target prediction and function annotation

The 3' untranslated region (3' UTR) of known human genes were retrieved from UCSC website. The most commonly used miRNA target scan program for mammalian, miRanda program [33], is used to predict the target genes of the known miRNAs and miRNA candidates. To demonstrate the biological significance of the target genes, several datasets of microarray were studied to obtain differentially expressed genes in normal and cancer tissues. The results are shown together with the target gene annotations.

The expression profiles of the miRNAs are included in the analyses to reveal the regulatory roles of the miRNAs. The expression data of normal liver tissue miRNAs were obtained from the RNA Project at Rockefeller [34]. The expression of known miRNAs retrieved from users' clone data are compared with that in normal liver using Fisher's exact test to find differentially expressed miRNAs in users' RNA clone and normal liver tissues.

To further view expression profiles of known miRNAs in liver, public available expression datasets of miRNA in liver were collected and processed. Different types of miRNA expression data, such as SAGE, microarray and bead-base array, were processed with different methods. For SAGE based data, Tags per million (TPM) of a miRNA, regarded as the expression level of this miRNA, is calculated and shown in our web display. For bead-base array and microarray, on the other hand, the log ratio value (M value) is normally directly provided by the data source. In some cases, several M values were provided and the average values were computed in our study. The M value represents overexpression or underexpression level of each miRNA, and is shown in the web interface.

The expression data of the target genes of known miRNAs were also provided. The microarray datasets were processed with the siggenes module of the R package from Bioconductor [35] which implements the algorithm of Significance Analysis of Microarrays (SAM) proposed by Tusher et al [36], using a modified t statistics by scoring each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. For SAGE data-

Table 1: Liver miRNA expression datasets in miRAS

Provider	Dataset Type	Tissue Description
Rockefeller Univ.*	SAGE	Mouse liver tissue
Shingara, J., et al. **	Microarray	Mouse liver tissue
Farh, K.K., et al.**	Microarray	Mouse liver tissue
Thomson, J.M., et al.***	Microarray	Mouse liver tissue
Lu, J., et al.***	Bead-base array	Human normal liver
Wienholds, E., et al.***	Microarray	Zebrafish tissues

* Datasets from the Rockefeller University RNA project [34].
** Datasets available from supplements of article, references [41] and [42], respectively.
*** Datasets available from Gene Expression Omnibus (GEO) [43] with accession number of GSE1635, GSE2564, and GSE2628, respectively.

Table 2: Liver gene expression datasets in miRAS

Dataset Type	Dataset Name	Tissue Description
Microarray:		
	Hep3B vs Universal_control*	Hepatoma cell line
	SNU_387 vs Universal_control*	Hepatoma cell line
	HepG2 vs Universal_control*	Hepatoma cell line
	PLC_PRF_5 vs Universal_control*	Hepatoma cell line
	SNU_182 vs Universal_control*	Hepatoma cell line
	SNU_354 vs Universal_control*	Hepatoma cell line
	SNU_368 vs Universal_control*	Hepatoma cell line
	SNU_449 vs Universal_control*	Hepatoma cell line
	SNU_475 vs Universal_control*	Hepatoma cell line
SAGE:		
	SAGE_Liver_cholangiocarcinoma_B_K1**	Liver
	SAGE_Liver_cholangiocarcinoma_B_K2D**	Liver
	SAGE_Liver_cholangiocarcinoma_CL_K3**	Liver cell line
	SAGE_Liver_cholangiocarcinoma_CL_K4**	Liver cell line
	SAGE_Liver_normal_B_1**	Normal bulk liver

* Datasets from GEO [43], GEO accession GSM81024- GSM81059
** Datasets from SAGE Genie [44] of CGAP, names are for library.

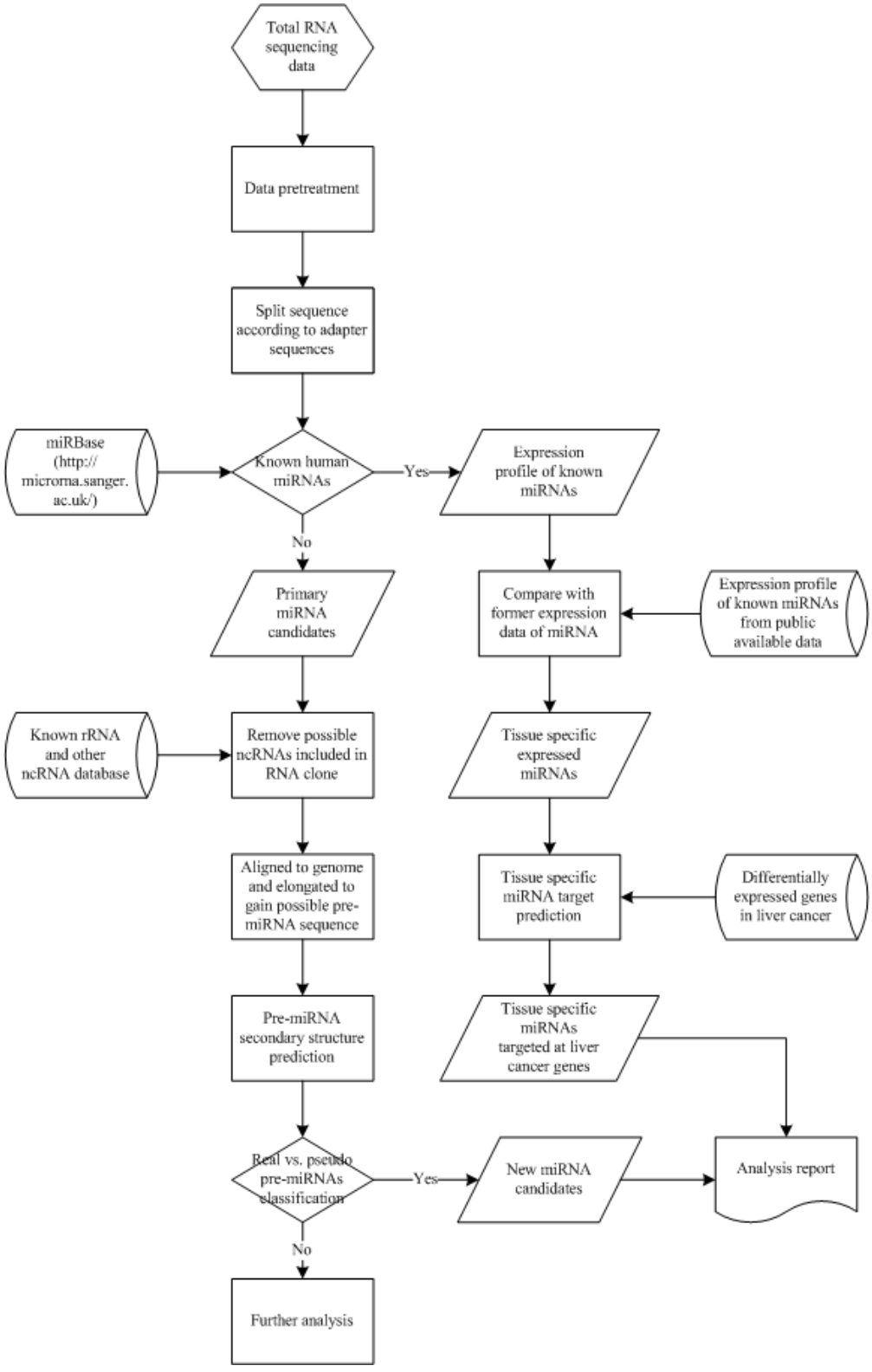

Figure 1
Work flow of miRAS.

sets, SAGE tags were mapped to known genes with the data provided by NCBI ftp site [37]. The differential expression between cancerous cell and normal cell was analyzed using Fisher's exact test [38-40], implemented with the sagenhaft module of R. For both types, M values were computed to represent the differential expression of the genes for cancerous vs. normal tissues.

The expression datasets of known miRNAs and their target genes included in miRAS are listed in Table 1 and Table 2.

Result retrieval

Through the result view page, the base-calling result, known mRNAs and other ncRNAs, possible miRNA candidates together with the secondary structures, target prediction result and their expressions in known datasets are returned. For users to view result in more convenience, different data formats are provided, such as raw data file, text file and Microsoft excel.

Competing interests

The authors declare that they have no competing interests.

Availability

miRAS is freely available on the web [21] suitable for most graphical web browser. User registered with an email address will be alerted the status of the submitted jobs. The system is also anonymously accessible.

Authors' contributions

FT proposed the analysis pipeline and completed the main part of this research. HZ implemented the CGI scripts and the system testing. SC and XZ helped in the webpage writing and the database for gene annotation. XZ also created a database for the processing of the SAGE data. YX, YW helped to review the manuscripts. XL supervised the design and implementation of the system, and advised on the manuscript preparation.

Acknowledgements

We thank our collaborator, Professor Jingde Zhu of Shanghai Cancer Institute, for providing the idea of building this analysis platform and target gene annotation for biological significance use. This project is supported by Trans-Century Training Programe Foundation for the Talents by the Chinese Ministry of Education, and by Project "BOD: A Grid-based Bioinformatics Computational Pipeline System" supported by National Natural Science Foundation of China (No. 90412018).

References

1. Bartel DP: **MicroRNAs: genomics, biogenesis, mechanism, and function.** *Cell* 2004, **116(2):**281-297.
2. Calin GA, Sevignani C, Dumitru CD, Hyslop T, Noch E, Yendamuri S, Shimizu M, Rattan S, Bullrich F, Negrini M, Croce CM: **Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers.** *Proc Natl Acad Sci U S A* 2004, **101(9):**2999-3004.
3. Gregory RI, Shiekhattar R: **MicroRNA biogenesis and cancer.** *Cancer Res* 2005, **65(9):**3509-3512.
4. Esquela-Kerscher A, Slack FJ: **Oncomirs - microRNAs with a role in cancer.** *Nat Rev Cancer* 2006, **6(4):**259-269.
5. Hammond SM: **MicroRNAs as oncogenes.** *Curr Opin Genet Dev* 2006, **16(1):**4-9.
6. Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, Tomida S, Yatabe Y, Kawahara K, Sekido Y, Takahashi T: **A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation.** *Cancer Res* 2005, **65(21):**9628-9632.
7. Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F, Visone R, Iorio M, Roldo C, Ferracin M, Prueitt RL, Yanaihara N, Lanza G, Scarpa A, Vecchione A, Negrini M, Harris CC, Croce CM: **A microRNA expression signature of human solid tumors defines cancer gene targets.** *Proc Natl Acad Sci U S A* 2006, **103(7):**2257-2261.
8. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR: **MicroRNA expression profiles classify human cancers.** *Nature* 2005, **435(7043):**834-838.
9. Iorio MV, Ferracin M, Liu CG, Veronese A, Spizzo R, Sabbioni S, Magri E, Pedriali M, Fabbri M, Campiglio M, Menard S, Palazzo JP, Rosenberg A, Musiani P, Volinia S, Nenci I, Calin GA, Querzoli P, Negrini M, Croce CM: **MicroRNA gene expression deregulation in human breast cancer.** *Cancer Res* 2005, **65(16):**7065-7070.
10. Yanaihara N, Caplen N, Bowman E, Seike M, Kumamoto K, Yi M, Stephens RM, Okamoto A, Yokota J, Tanaka T, Calin GA, Liu CG, Croce CM, Harris CC: **Unique microRNA molecular profiles in lung cancer diagnosis and prognosis.** *Cancer Cell* 2006, **9(3):**189-198.
11. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, Rassenti L, Kipps T, Negrini M, Bullrich F, Croce CM: **Frequent deletions and down-regulation of microRNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.** *Proc Natl Acad Sci U S A* 2002, **99(24):**15524-15529.
12. Calin GA, Liu CG, Sevignani C, Ferracin M, Felli N, Dumitru CD, Shimizu M, Cimmino A, Zupo S, Dono M, Dell'Aquila ML, Alder H, Rassenti L, Kipps TJ, Bullrich F, Negrini M, Croce CM: **MicroRNA profiling reveals distinct signatures in B cell chronic lymphocytic leukemias.** *Proc Natl Acad Sci U S A* 2004, **101(32):**11755-11760.
13. Kim VN, Nam JW: **Genomics of microRNA.** *Trends Genet* 2006, **22(3):**165-173.
14. Liu CG, Calin GA, Meloon B, Gamliel N, Sevignani C, Ferracin M, Dumitru CD, Shimizu M, Zupo S, Dono M, Alder H, Bullrich F, Negrini M, Croce CM: **An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues.** *Proc Natl Acad Sci U S A* 2004, **101(26):**9740-9744.
15. Hammond SM: **microRNA detection comes of age.** *Nat Methods* 2006, **3(1):**12-13.
16. Cummins JM, He Y, Leary RJ, Pagliarini R, Diaz LA Jr., Sjoblom T, Barad O, Bentwich Z, Szafranska AE, Labourier E, Raymond CK, Roberts BS, Juhl H, Kinzler KW, Vogelstein B, Velculescu VE: **The colorectal microRNAome.** *Proc Natl Acad Sci U S A* 2006, **103(10):**3687-3692.
17. Wang SM: **Understanding SAGE data.** *Trends Genet* 2007, **23(1):**42-50.
18. Tuteja R, Tuteja N: **Serial analysis of gene expression (SAGE): unraveling the bioinformatics tools.** *Bioessays* 2004, **26(8):**916-922.
19. Lash AE, Tolstoshev CM, Wagner L, Schuler GD, Strausberg RL, Riggins GJ, Altschul SF: **SAGEmap: a public gene expression resource.** *Genome Res* 2000, **10(7):**1051-1060.
20. Boon K, Osorio EC, Greenhut SF, Schaefer CF, Shoemaker J, Polyak K, Morin PJ, Buetow KH, Strausberg RL, De Souza SJ, Riggins GJ: **An anatomy of normal and malignant gene expression.** *Proc Natl Acad Sci U S A* 2002, **99(17):**11287-11292.
21. **miRAS system** [http://e-science.tsinghua.edu.cn/miras]
22. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome research* 1998, **8(3):**186-194.
23. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome research* 1998, **8(3):**175-185.
24. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ: **miRBase: microRNA sequences, targets and gene nomenclature.** *Nucleic Acids Res* 2006, **34(Database issue):**D140-4.
25. Griffiths-Jones S: **The microRNA Registry.** *Nucleic Acids Res* 2004, **32(Database issue):**D109-11.
26. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM: **The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis.** *Nucleic Acids Res* 2005, **33(Database issue):**D294-6.
27. Liu C, Bai B, Skogerbo G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R: **NONCODE: an integrated knowledge database of noncoding RNAs.** *Nucleic Acids Res* 2005, **33(Database issue):**D112-5.

28. Szymanski M, Erdmann VA, Barciszewski J: **Noncoding regulatory RNAs database.** *Nucleic Acids Res* 2003, **31(1):**429-431.
29. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31(13):**3406-3415.
30. I. L. Hofacker WF P. F. Stadler, L. S. Bonhoeffer, M. Tacker and P. Schuster: **Fast folding and comparison of RNA secondary structures.** In *Monatshefte für Chemie Volume 125*. Issue Volume 125, Number 2 / February, 1994 Springer Wien; 1994:167-188.
31. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, Matzke M, Ruvkun G, Tuschl T: **A uniform system for microRNA annotation.** *Rna* 2003, **9(3):**277-279.
32. Xue C, Li F, He T, Liu GP, Li Y, Zhang X: **Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine.** *BMC Bioinformatics* 2005, **6:**310.
33. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS: **Human MicroRNA targets.** *PLoS Biol* 2004, **2(11):**e363.
34. **RNA Project at Rockefeller** [http://www.rockefeller.edu/labheads/tuschl/mirna.html]
35. **Bioconductor** [http://www.bioconductor.org/]
36. Tusher VG, Tibshirani R, Chu G: **Significance analysis of microarrays applied to the ionizing radiation response.** *Proc Natl Acad Sci U S A* 2001, **98(9):**5116-5121.
37. **NCBI ftp site** [ftp://ftp.ncbi.nih.gov/]
38. Fisher L, van Belle G: **Biostatistics:A Methodology for the Health Sciences.** New York , John Wiley and Sons,; 1993:991.
39. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7(10):**986-995.
40. Man MZ, Wang X, Wang Y: **POWER_SAGE: comparing statistical tests for SAGE experiments.** *Bioinformatics* 2000, **16(11):**953-959.
41. Shingara J, Keiger K, Shelton J, Laosinchai-Wolf W, Powers P, Conrad R, Brown D, Labourier E: **An optimized isolation and labeling platform for accurate microRNA expression profiling.** *Rna* 2005, **11(9):**1461-1470.
42. Farh KK, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP: **The widespread impact of mammalian MicroRNAs on mRNA repression and evolution.** *Science* 2005, **310(5755):**1817-1821.
43. **Gene Expression Omnibus (GEO)** [http://www.ncbi.nlm.nih.gov/geo/]
44. **SAGE Genie** [http://cgap.nci.nih.gov/SAGE]