Methods in Molecular Biology (2016) 1375: 75–89
DOI 10.1007/7651_2015_237

Published online: 11 April 2015

MetaMirClust: Discovery and Exploration of Evolutionarily Conserved miRNA Clusters

Wen-Ching Chan and Wen-chang Lin

Abstract

Recent emerging studies suggest that a substantial fraction of microRNA (miRNA) genes is likely to form clusters in terms of evolutionary conservation and biological implications, posing a significant challenge for the research community and shifting the bottleneck of scientific discovery from miRNA singletons to miRNA clusters. In addition, the advance in molecular sequencing technique such as next-generation sequencing (NGS) has facilitated researchers to comprehensively characterize miRNAs with low abundance on genome-wide scale in multiple species. Taken together, a large scale, cross-species survey of grouped miRNAs based on genomic location would be valuable for investigating their biological functions and regulations in an evolutionary perspective. In the present chapter, we describe the application of effective and efficient bioinformatics tools on the identification of clustered miRNAs and illustrate how to use the recently developed Web-based database, MetaMirClust (http://fgfr.ibms.sinic.aedu.tw/MetaMirClust) to discover evolutionarily conserved pattern of miRNA clusters across metazoans.

Keywords: MetaMirClust, microRNA cluster, Data mining, Synteny

1 Introduction

MicroRNAs (miRNAs) are, of 21–23 nucleotides (nt) long in their mature forms, a recently identified class of endogenous small noncoding RNA molecules, which play important roles in gene regulation via the RNA interference pathway (1–4). In 1993, when the first miRNA *lin-4* was identified in *Caenorhabditis elegans*, the negative regulation pair between *lin-4* and its target *lin-14* was thought as an individual case (5). As a result, miRNAs have not gained the attention of researchers until a second similar system of *let-7* was observed (6), and then its homologous transcripts were extensively investigated in animal and plant genomes. In these two decades, a considerable body of evidence suggests that miRNAs play important gene-regulatory roles related to organism development, cell differentiation, and tumor progression and oncogenesis (7–11). Currently, newly discovered miRNA genes either by experimental or computational approaches have steadily increased as evident by the amount of records in the miRBase registry (12) and other resources (13, 14). In recent years, many studies have attempted to gain insights into the biogenesis, expression, targeting, and evolution of individual miRNA gene in different species. Some well-studied examples in human are, for instance, *mir-34b* and *mir-129* which serve as tumor-suppressor miRNAs connected to DNA methylation-associated silencing in gastric cancer (10); *mir-196a* is overexpressed in primary gastric cancer tissues compared to adjacent normal ones (9); three individual loci of *mir-9* are simultaneously hypermethylated in gastric cancer and are likely to serve as tumor suppressive miRNAs (8). Correspondingly, a substantial amount of literature has demonstrated miRNAs as crucial negative regulators in diverse physiological and developmental processes at the posttranscriptional level (15).

Up to date, a handful of miRNA clusters have been reported in animal genomes. To the best of our knowledge, Altuvia et al. was the first group that identified conserved regions of miRNA clusters systematically (16). Then, Yu et al. adopted the same method to enlarge the extent of conserved miRNA cluster (17), and thus checked the expression profile of identified human miRNA clusters. Accumulating studies have illustrated that clustered miRNA genes located on polycistronic transcripts might be expressed at similar levels and coordinately involve in an intricate regulatory network. These miRNA clusters are usually derived from polycistrons within the length from few hundred nucleotides to almost million base pairs (18–21). For instance, *mir-17* cluster and its paralogous clusters are one of the well-studied cases. In 2004, Tanzer et al. have tried to reconstruct the phylogenetic evolution of *mir-17* cluster family mainly in nine metazoan genomes and have revealed at least three paralogous clusters related to the *mir-17* cluster family, which are *mir-17-92*, *mir-106-92*, and *mir-106-25*, and governed by tandem duplications (22).

A growing range of studies has further demonstrated that the aberrant expression of miRNAs in cluster families plays an important role in cancer oncogenesis and metastasis (23–25). In addition to the known function of *mir-92a* as negative regulator of angiogenesis, an overexpression pattern of the *mir-17-92* cluster (13q31.3) comprising seven miRNAs has been discovered in 19 lung cancer cell lines (26). In renal cell carcinoma (RCC), the restoration of the downregulated *mir-143/145* cluster (5q32) in RCC cells revealed significant inhibition of cancer cell proliferation and invasion via a putative target gene, hexokinase-2 (HK2) (27). In bladder cancer (BC), five downregulated clusters: *mir-1/133a* (18q11.2 and 20q13.33), *mir-206/133b* (p12.2), *let-7c/mir-99a* (21q21.1), *mir-143/145* (5q32), and *mir-195/497* (17p13.1), were identified from 950 candidates by the genome-wide miRNA expression signature analysis, and the following transfection assay of *mir-195/497* into BC cell lines has confirmed their function as tumor suppressors in BC (25). It is believed that miRNAs in clusters might represent putative bifunctional regulators, of which miRNAs in high expression level can act as oncogenes by repressing tumor suppressors, and when in low level they can turn over to behave as tumor suppressors through a negative regulation of oncogenes (28). Although the entire regulatory mechanisms of clustered miRNA genes remain largely uncharacterized, it is likely that these miRNA clusters may function more efficiently in a complicated miRNA-mediated network than individual miRNA alone (29). Therefore, identification of evolutionary conserved miRNA clusters is an important first step for the research society toward elucidating miRNA-cluster-mediated pathways in cancer research and might provide new insights into the potential miRNA-based therapeutics for cancer.

Many resources were developed to investigate miRNA genes. However, only a handful of resources dedicate to an efficient and extensive investigation of miRNA clusters (20, 21). Generally, miRNA clusters were arbitrarily defined by a fixed distance (e.g., 10 Kb) (12), and only few studies systematically investigating the conservation patterns of clustered miRNA genes across metazoan species (20). Here, we illustrate the synergistic potential of MetaMirClust and miRBase for exploring miRNA clusters conserved across species in evolution.

The remainder of the chapter is organized as follows. First, the Materials section highlights the technical prerequisites for the identification of miRNA clusters used in MetaMirClust; second, we give an overview of available databases that enlarge the scope of miRNA genes; third, we introduce how to identify miRNA clusters (MirClust) in different maximum inter-miRNA distances (MIDs) as well as a simple case study of using and browsing MirClust; fourth, we outline the use of MetaMirClust for exploring metazoan conserved miRNA clusters and their hierarchically evolutionary structure; fifth, we describe an advanced case study that uses bioinformatics tools and additional annotation files to uncover the synteny regions flanking miRNA clusters between human and mouse. Finally, in the Notes section, we briefly comment on practical issues and highlight potential pitfalls of the methods that are outlined in this chapter.

2 Materials

MetaMirClust is a Web-based database and can be browsed via a user-friendly interface implemented according to the protocols of HyperText Markup Language (HTML) and Cascading Style Sheets (CSS). For general users who focus on browsing data in MetaMirClust, the community user can easily access it using a computer with an Internet network. For instance, it can be a desktop computer running Microsoft Windows, an Apple computer running Mac OS, or a LINUX platform. A few commonly used browsers include (1) Mozilla Firefox (http://www.firefox.com/), (2) Microsoft Internet Explorer (http://www.microsoft.com/ie/), (3) Apple Safari (http://www.apple.com/de/safari/), and (4) Google Chrome (https://www.google.com/intl/en/chrome/).

For advanced users who want to re-perform the whole analysis procedure and/or follow-up analyses (i.e., the identification of synteny regions between human and mouse), beyond the essential Web browser, it is recommended to install an advanced text editor, e.g., Sublime Text 2/3 (http://www.sublimetext.com/) or Programmer's Notepad (http://www.pnotepad.org/), which can effectively and efficiently facilitate scripting jobs and which manipulates large files (e.g., table-delimited BED files with gene models from UCSC Table Browser) and/or data format conversions. When dealing with BED format files, BEDTools 2 (https://github.com/arq5x/bedtools2) is one of fundamental tools, which efficiently manages the operations like merging, intersecting, and/or subtracting between two BED files. In addition, to build a SQL-like environment to contain data downloaded from public resources like miRBase or to store intermediate results generated through the pipeline, MySQL is one of the best choices for a fast, multi-threads/users and robust database management system. In MetaMirClust, we introduced a data mining approach, i.e., FP-growth (30, 31), to efficiently discover highly conserved sets of miRNA genes upon miRNA clusters (MirClust). The implementation version of FP-growth algorithm by Borgelt is available to download (http://www.borgelt.net/fpgrowth.html). Similarly, the final results after the mining procedure are restored into MySQL database for querying and browsing via the Web-based interface. Finally, for visualization, it is also useful to install the perl models like GD (http://search.cpan.org/~lds/GD/) as well as the R statistics software (http://www.r-project.org/) to present results in image files for visual inspection.

3 Methods

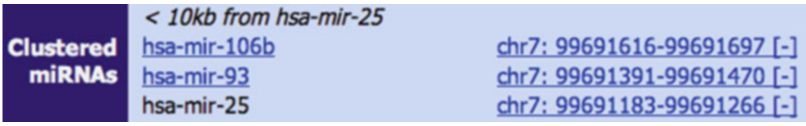
3.1 Homology Search of miRNA Genes

A comprehensive understanding of miRNA clusters will require an extensive survey of the coverage of miRNA genes in genomes. Previously, miRNA genes were identified through cloning and sequencing of small-RNA libraries. However, miRNA genes could be overlooked due to low expression levels. In this decade, the ever-growing data adopted next-generation sequencing (NGS) technique to identify miRNA genes has been incorporated into public databases like miRBase. Since those studies were mainly focusing on a small set of species, it is still necessary to conduct an extensive homology search based on known miRNA genes collected in miRBase to enlarge the scope of miRNA genes across mammals. The current version of MetaMirClust has been performed based on known miRNA genes reported in miRBase (Release 16: Sept

2010) and predicted homologous miRNA genes in ZooMir (http://insr.ibms.sinica.edu.tw/zoomir/) (14). The data of the ZooMir version used in the current MetaMirClust are dumped from MySQL and can be downloaded (http://insr.ibms.sinica.edu.tw/ZooMir/ZooMir.Candidates_3.tar.bz2). Using the characteristics of sequence- and structure-conservation of miRNA genes, additional 14,989 homologous precursor miRNA candidates in 56 genomes have been identified according to 11,839 animal miRNA entries reported in miRBase 16.0. In addition, we classified miRNA genes by reassigning miRNA classes based on the sequence similarity with same prefix of their entry names without considering species abbreviations used in miRBase.

3.2 Identification of miRNA Clusters (MirClust)

Recent studies have revealed that the clustering propensity of miRNA genes is higher than previously evaluated and they usually occur on polycistronic transcripts (17, 32–36). To investigate clustered miRNA genes derived from the same polycistronic transcript, researchers usually adopt adjacent miRNA genes located on the same strand to form miRNA clusters. Two or more consecutive miRNA genes on the same strand of individual chromosome are considered to form a cluster according to their adjacent distance. In miRBase, 10 Kb is used to report clustered miRNAs when users browse an individual miRNA gene. Take *hsa-mir-25* (chr7:99,691,183-99,691,266:-) as example, miRBase will display *hsa-mir-93* (chr7:99,691,391-99,691,470:-) and *hsa-mir-106b* (chr7:99,691,616-99,691,697:-) as adjacent miRNA genes within 10 Kb as shown in Fig. 1. As a result, using different adjacent distance might result in a different data set of miRNA clusters. Meanwhile, the clustered miRNAs reported in miRBase are lack of evolutionary conservation across species. Four different maximum inter-miRNA distances (MIDs); 1 Kb, 3 Kb, 10 Kb, and 50 Kb, were commonly used to identify clustered miRNA genes (MirClust). To illustrate the procedure of identification of miRNA clusters (MirClust), we prepared two BED file composed of human (hg19) precursor/mature miRNA genes (reported in miRBase v.16 or ZooMir) (http://fgfr.ibms.sinica.edu.tw/MetaMirClust/data/pre.mir.bed; http://fgfr.ibms.sinica.edu.tw/MetaMirClust/data/mat.mir.bed) as a sample data set for readers to identify miRNA clusters (MirClust) in human. In addition, the BED file of individual mature miRNA genes was prepared for the retrieval of miRNA



Clustered miRNAs	< 10kb from hsa-mir-25	
	hsa-mir-106b	chr7: 99691616-99691697 [-]
	hsa-mir-93	chr7: 99691391-99691470 [-]
	hsa-mir-25	chr7: 99691183-99691266 [-]

Fig. 1 The *hsa-mir-25-106b* cluster reported in miRBase. By the default MID of 10 Kb used in miRBase, this snapshot figure shows two adjacent miRNA genes, *hsa-mir-106b* and *hsa-mir-93*, when querying *hsa-mir-25*

clusters with their corresponding mature miRNAs. The individual processes were listed as follows.

1. Sort precursor miRNA genes:

   ```
   sort -k1,1 -k2,2n -k6,6 pre.mir.bed > pre.mir.sort.bed
   ```

2. Group miRNA genes to form miRNA clusters based on user-defined MID

   ```
   bedtools merge -s -d 10000 -c 4,6,4 -o collapse,distinct,count -i pre.mir.sort.bed > mir.clust.bed
   ```

3. Remove singleton miRNA clusters

   ```
   awk 'BEGIN{OFS=FS="\t"}{if ($6 > 1) {print $0}}' mir.clust.be > mir.clust.filter.bed
   ```

4. (Optional) Retrieve mature miRNA genes for each miRNA cluster

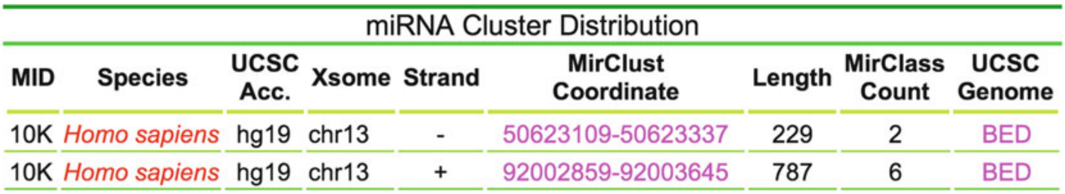
   ```
   bedtools intersect -wo -a mir.clust.filter.bed -b mat.mir.bed > mir.clust.mat.bed
   ```

The above command would create two intermediate files (i.e., pre.mir.sort.bed and mir.clust.bed) plus one output file for the final result of miRNA clusters (i.e., mir.clust.filter.bed). First, to prepare the sorted BED file as the input file of BEDTools in the following process, the human miRNA genes in the BED file were sorted according to their genomic location plus strand information. Subsequently, using the merge command in the BEDTools package, adjacent miRNA genes were grouped according to the user-defined MID (here, 10 Kb). The grouped miRNAs passing the third step by filtering singleton miRNA clusters will create miRNA clusters in human in this sample example. Correspondingly, the whole procedure can be achieved by piping into one command line: ***bedtools merge -s -d 10000 -c 4,6,4 -o collapse,distinct,count -i < (sort -k1,1 -k2,2n -k6,6 pre.mir.bed) | awk 'BEGIN{OFS=FS="\t"}{if ($6 > 1) {print $0}}' > mir.clust.filter.bed***.

By comparing miRNA clusters discovered in a short MID to those in a longer one, three scenarios are discovered: (1) forming a new miRNA cluster by merging singleton miRNA genes, (2) enlarging a small miRNA cluster by recruiting singleton miRNA genes, and (3) producing a large miRNA cluster by merging at least two small miRNA clusters. According to our previous observation (20), when considering a long MID, these newly involved clustered miRNA genes are apt to generate new miRNA clusters instead of enlarging miRNA clusters in a short MID. It is suggestive of that miRNA genes are prone to form clusters, and those miRNA clusters are separately located far away from each other. Table 1 shows the distributions of numbers of miRNA clusters (MirClust) identified using four different MID in nine representative species, including *Caenorhabditis elegans* (worm, ce6), *Drosophila melanogaster* (fly, dm3), *Danio rerio* (zebrafish, danRer6), *Gallus gallus* (chicken,

Table 1
Distributions of numbers of identified miRNA clusters in nine representative species

		MID			
Species	**UCSC accession**	**1 Kb**	**3 Kb**	**10 Kb**	**50 Kb**
Caenorhabditis elegans	ce6	13	18	26	38
Drosophila melanogaster	dm3	19	18	21	33
Danio rerio	danRer6	38	55	61	73
Gallus gallus	galGal3	22	41	54	72
Canis familiaris	canFam2	50	49	57	74
Bos taurus	bosTau4	56	54	61	81
Mus musculus	mm9	78	65	69	84
Rattus norvegicus	rn4	57	60	63	70
Homo sapiens	hg19	66	74	79	100

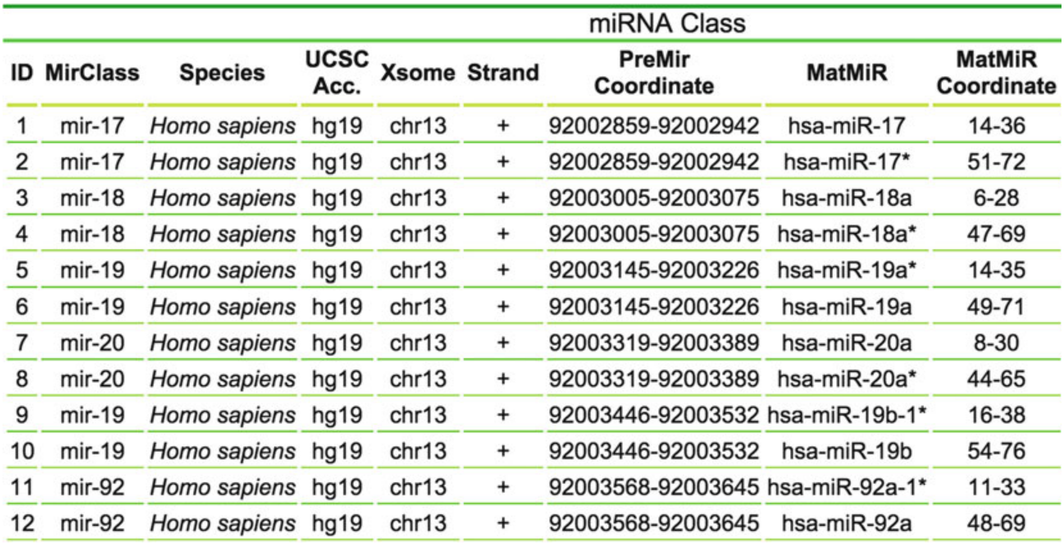


miRNA Cluster Distribution

MID	Species	UCSC Acc.	Xsome	Strand	MirClust Coordinate	Length	MirClass Count	UCSC Genome
10K	Homo sapiens	hg19	chr13	-	50623109-50623337	229	2	BED
10K	Homo sapiens	hg19	chr13	+	92002859-92003645	787	6	BED

Fig. 2 Two miRNA clusters identified on chromosome 13 in human. Based on miRNA genes identified in miRBase and ZooMir and the use of MID of 10 Kb, two miRNA clusters can be revealed on chromosome 13. One is *mir-15a/16* (13q14.2) in the length of 229 nt on the plus strand and the other is *mir-17-92* (13q31.3) in the length of 787 nt on the minus strand

galGal3), *Canis familiaris* (dog, canFam2), *Bos taurus* (cow, bosTau4), *Mus musculus* (mouse, mm9), *Rattus norvegicus* (rat, rn4), and *Homo sapiens* (human, hg19). According to the sample example, Fig. 2 lists two miRNA clusters (MirClust) identified on chromosome 13 in human according to the MID of 10 Kb, which are *mir-16-1/15a* (13q14.2) and *mir-17-92* (13q31.3). The community users can retrieve the detailed information of individual mature miRNA genes for each miRNA clusters through the forth, optional command listed above. Correspondingly, through our Web-based interface (http://fgfr.ibms.sinica.edu.tw/MetaMirClust/MirClustStat.php), the community users can browse related information of *mir-17-92* (13q31.3) as shown in Fig. 3. The links to external browsers like UCSC Genome Browser (https://genome.ucsc.edu/) are provided to obtain more information about miRNA clusters (e.g., conservation levels and transcriptions in RefSeq or GenBank). Figure 4 shows several default tracks in the genomic location flanking *mir-17-92* (13q31.3) in UCSC Genome Browser.



						miRNA Class		
ID	MirClass	Species	UCSC Acc.	Xsome	Strand	PreMir Coordinate	MatMiR	MatMiR Coordinate
1	mir-17	*Homo sapiens*	hg19	chr13	+	92002859-92002942	hsa-miR-17	14-36
2	mir-17	*Homo sapiens*	hg19	chr13	+	92002859-92002942	hsa-miR-17*	51-72
3	mir-18	*Homo sapiens*	hg19	chr13	+	92003005-92003075	hsa-miR-18a	6-28
4	mir-18	*Homo sapiens*	hg19	chr13	+	92003005-92003075	hsa-miR-18a*	47-69
5	mir-19	*Homo sapiens*	hg19	chr13	+	92003145-92003226	hsa-miR-19a*	14-35
6	mir-19	*Homo sapiens*	hg19	chr13	+	92003145-92003226	hsa-miR-19a	49-71
7	mir-20	*Homo sapiens*	hg19	chr13	+	92003319-92003389	hsa-miR-20a	8-30
8	mir-20	*Homo sapiens*	hg19	chr13	+	92003319-92003389	hsa-miR-20a*	44-65
9	mir-19	*Homo sapiens*	hg19	chr13	+	92003446-92003532	hsa-miR-19b-1*	16-38
10	mir-19	*Homo sapiens*	hg19	chr13	+	92003446-92003532	hsa-miR-19b	54-76
11	mir-92	*Homo sapiens*	hg19	chr13	+	92003568-92003645	hsa-miR-92a-1*	11-33
12	mir-92	*Homo sapiens*	hg19	chr13	+	92003568-92003645	hsa-miR-92a	48-69

Fig. 3 The mature miRNA genes located in *mir-17-92*. The *mir-17-92* (13q31.3) cluster located on chromosome 13 consists of six precursor miRNA genes and will encode 12 mature miRNA genes

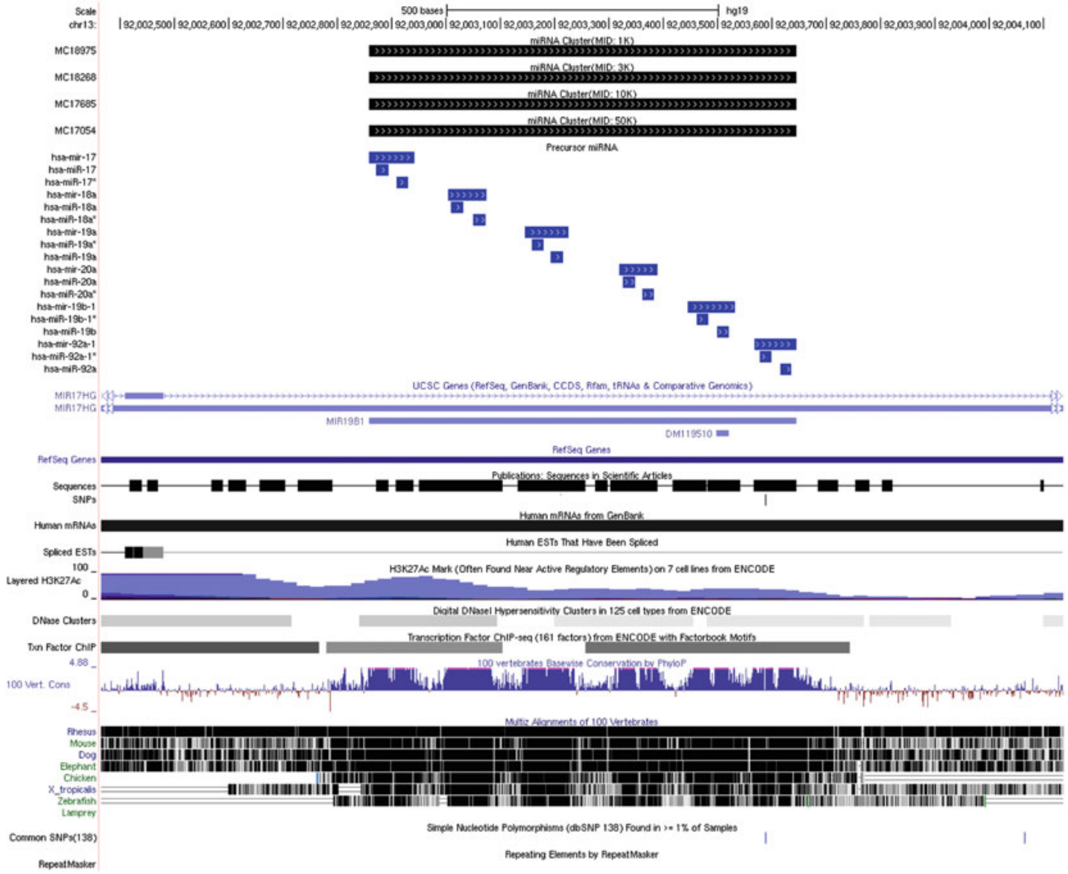


Fig. 4 MetaMirClust data of *mir-17-92* shown in UCSC Genome Browser. Viewing the cluster information of human *mir-17-92* (13q31.3) cluster using public genome browser like UCSC Genome Browser, additional pieces of evidence such as transcriptional regions, histone modifications, and conservation level and so on can facilitate users to gain more insights of miRNA clusters of interest

3.3 Discovery of Metazoan miRNA Clusters (MetaMirClust) by FP-Growth Algorithm

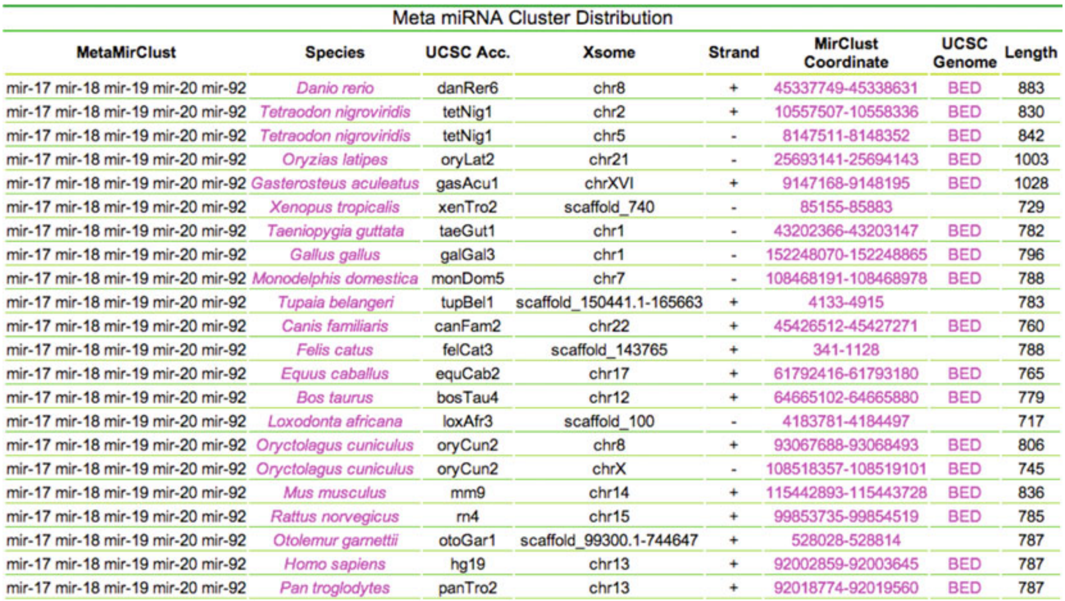
Most previous works only focused on studying the evolutionary and functional implications of limited specific miRNA clusters among a few species. No systematic and efficient approach has been performed before MetaMirClust to analyze the conservation pattern of miRNA clusters on global-wide scale. To interrogate the conservation level of the clusters of miRNA genes in large numbers of metazoan genomes, we adopted a data mining approach to discover the conserved co-occurrence modules of miRNA genes upon miRNA clusters identified under the same MID. Filtering singleton miRNA clusters identified in MirClust as mentioned in the previous procedure, we conducted the analysis by utilizing the FP-growth algorithm implemented by Borgelt (http://www.borgelt.net/fpgrowth.html) to detect the conserved co-occurrence sets of miRNA genes in terms of miRNA clusters defined within the same MID. These frequent co-occurrence sets present highly conserved combinations of miRNA genes through miRNA clusters in metazoan species, which are defined as metazoan miRNA clusters. Based on nine representative species same as listed in Table 1, we prepared an aggregate file (http://fgfr.ibms.sinica.edu.tw/MetaMirClust/data/nine.mir.clust.csv) consisting of all miRNA clusters using the previous procedure to identify MirClust. The following command can be used to discover co-occurred miRNA genes across selected species.

1. Discover co-occurred miRNA genes across species

```
fpgrowth -s-7 -q0 nine.mir.clust.csv nine.meta.mir.clust.csv
```

According to the output result (i.e., nine.meta.mir.clust.csv), there are 84 evolutionarily conserved miRNA clusters (MetaMirClust) identified in at least seven out of nine representative species. Among those evolutionarily conserved miRNA clusters, *mir-17-92* (13q31.3) is the largest group containing five miRNA classes with six precursor miRNA genes. Figure 5 shows the conservation pattern of *mir-17-92* (13q31.3) in MetaMirClust. The length of the *mir-17-92* (13q31.3) cluster varies from 717 (*Loxodonta africana*) to 1,028 (*Gasterosteus aculeatus*) nucleotides (nt) in 20 metazoan genomes, which confirmed the estimation of the *mir-17* cluster length as 1 kb reported previously.

In MetaMirClust, to investigate the recruitment process between evolutionarily conserved miRNA clusters, we also reconstructed the hierarchical structure using the sets of co-occurred miRNA genes. The community users can directly select one of evolutionarily conserved miRNA clusters of interest from the MetaMirClust list (http://fgfr.ibms.sinica.edu.tw/MetaMirClust/MetaMirClustStat.php) or select one of miRNA classes from the search page in MetaMirClust (http://fgfr.ibms.sinica.edu.tw/MetaMirClust/MetaMirClustSearch.php) to obtain the hierarchical information involving the selected miRNA cluster and the



Meta miRNA Cluster Distribution

MetaMirClust	Species	UCSC Acc.	Xsome	Strand	MirClust Coordinate	UCSC Genome	Length
mir-17 mir-18 mir-19 mir-20 mir-92	*Danio rerio*	danRer6	chr8	+	45337749-45338631	BED	883
mir-17 mir-18 mir-19 mir-20 mir-92	*Tetraodon nigroviridis*	tetNig1	chr2	+	10557507-10558336	BED	830
mir-17 mir-18 mir-19 mir-20 mir-92	*Tetraodon nigroviridis*	tetNig1	chr5	-	8147511-8148352	BED	842
mir-17 mir-18 mir-19 mir-20 mir-92	*Oryzias latipes*	oryLat2	chr21	-	25693141-25694143	BED	1003
mir-17 mir-18 mir-19 mir-20 mir-92	*Gasterosteus aculeatus*	gasAcu1	chrXVI	+	9147168-9148195	BED	1028
mir-17 mir-18 mir-19 mir-20 mir-92	*Xenopus tropicalis*	xenTro2	scaffold_740	-	85155-85883		729
mir-17 mir-18 mir-19 mir-20 mir-92	*Taeniopygia guttata*	taeGut1	chr1	-	43202366-43203147	BED	782
mir-17 mir-18 mir-19 mir-20 mir-92	*Gallus gallus*	galGal3	chr1	-	152248070-152248865	BED	796
mir-17 mir-18 mir-19 mir-20 mir-92	*Monodelphis domestica*	monDom5	chr7	-	108468191-108468978	BED	788
mir-17 mir-18 mir-19 mir-20 mir-92	*Tupaia belangeri*	tupBel1	scaffold_150441.1-165663	+	4133-4915		783
mir-17 mir-18 mir-19 mir-20 mir-92	*Canis familiaris*	canFam2	chr22	+	45426512-45427271	BED	760
mir-17 mir-18 mir-19 mir-20 mir-92	*Felis catus*	felCat3	scaffold_143765	+	341-1128		788
mir-17 mir-18 mir-19 mir-20 mir-92	*Equus caballus*	equCab2	chr17	+	61792416-61793180	BED	765
mir-17 mir-18 mir-19 mir-20 mir-92	*Bos taurus*	bosTau4	chr12	+	64665102-64665880	BED	779
mir-17 mir-18 mir-19 mir-20 mir-92	*Loxodonta africana*	loxAfr3	scaffold_100	-	4183781-4184497		717
mir-17 mir-18 mir-19 mir-20 mir-92	*Oryctolagus cuniculus*	oryCun2	chr8	+	93067688-93068493	BED	806
mir-17 mir-18 mir-19 mir-20 mir-92	*Oryctolagus cuniculus*	oryCun2	chrX	-	108518357-108519101	BED	745
mir-17 mir-18 mir-19 mir-20 mir-92	*Mus musculus*	mm9	chr14	+	115442893-115443728	BED	836
mir-17 mir-18 mir-19 mir-20 mir-92	*Rattus norvegicus*	rn4	chr15	+	99853735-99854519	BED	785
mir-17 mir-18 mir-19 mir-20 mir-92	*Otolemur garnettii*	otoGar1	scaffold_99300.1-744647	+	528028-528814		787
mir-17 mir-18 mir-19 mir-20 mir-92	*Homo sapiens*	hg19	chr13	+	92002859-92003645	BED	787
mir-17 mir-18 mir-19 mir-20 mir-92	*Pan troglodytes*	panTro2	chr13	+	92018774-92019560	BED	787

Fig. 5 The conservation pattern of *mir-17-92* across metazoan. The *mir-17-92* (13q31.3) cluster has been revealed to conserve across 20 species in terms of evolution in our data set and occurs 24 instances among these species

occurrence in each species under different MIDs. Take *mir-25* as example, the search result under the MID of 10 Kb is shown as Table 2 with all evolutionarily conserved miRNA clusters containing the target *mir-17* miRNA. For visualization, the drawing of conservation pattern upon genomes across species has been provided in MetaMirClust as shown in Fig. 6.

4 Notes

4.1 Data Preparation from Diverse Sources

In miRNA research, miRBase is the most critical repository, in which computational and experimental miRNA genes have been collected, and a searchable database. Recently, due to the advance in molecular sequencing technique like next-generation sequencing (NGS), miRBase have obtained ever-growing miRNA genes identified from the screening experiments (37). Currently, the miRBase database provides two major formats of archive files: raw-text and SQL-like files. The former includes dat and fa files in EMBL and fasta formats, respectively. They are easily for the community users to check the RNA sequences of precursor and mature miRNA genes. On the other hand, the SQL-like files dumped directly from miRBase contain more information, which is normalized and store into individual tables in terms of database management. For advanced users, the latter files will be more efficient to retrieve related data

Table 2
Hierarchical structure of different recruitment of *mir-25-106*

		SID																													
miRNA Cluster	Occurrence	*Danio rerio*	*Fugu rubripes*	*Tetraodon nigroviridis*	*Oryzias latipes*	*Gasterosteus aculeatus*	*Xenopus tropicalis*	*Anolis carolinensis*	*Taeniopygia guttata*	*Gallus gallus*	*Ornithorhynchus anatinus*	*Monodelphis domestica*	*Dasypus novemcinctus*	*Erinaceus europaeus*	*Echinops telfari*	*Sorex araneus*	*Tupaia belangeri*	*Canis familiaris*	*Felis catus*	*Equus caballus*	*Bos taurus*	*Loxodonta africana*	*Oryctolagus cuniculus*	*Cavia porcellus*	*Mus musculus*	*Rattus norvegicus*	*Callithrix jacchus*	*Otolemur garnettii*	*Macaca mulatta*	*Homo sapiens*	*Pan troglodytes*	*Pongo abelii*
mir-25 mir-93 mir-106	15													1			1	1	1	1	1		1	1	1	1	1	1	1	1	1	
mir-25 mir-93	20	1					2					1		1		1	1	1	1	1	1		1	1	1	1	1	1	1	1	1	
mir-93 mir-106	18													1	1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
mir-25 mir-106	15													1			1	1	1	1	1		1	1	1	1	1	1	1	1	1	

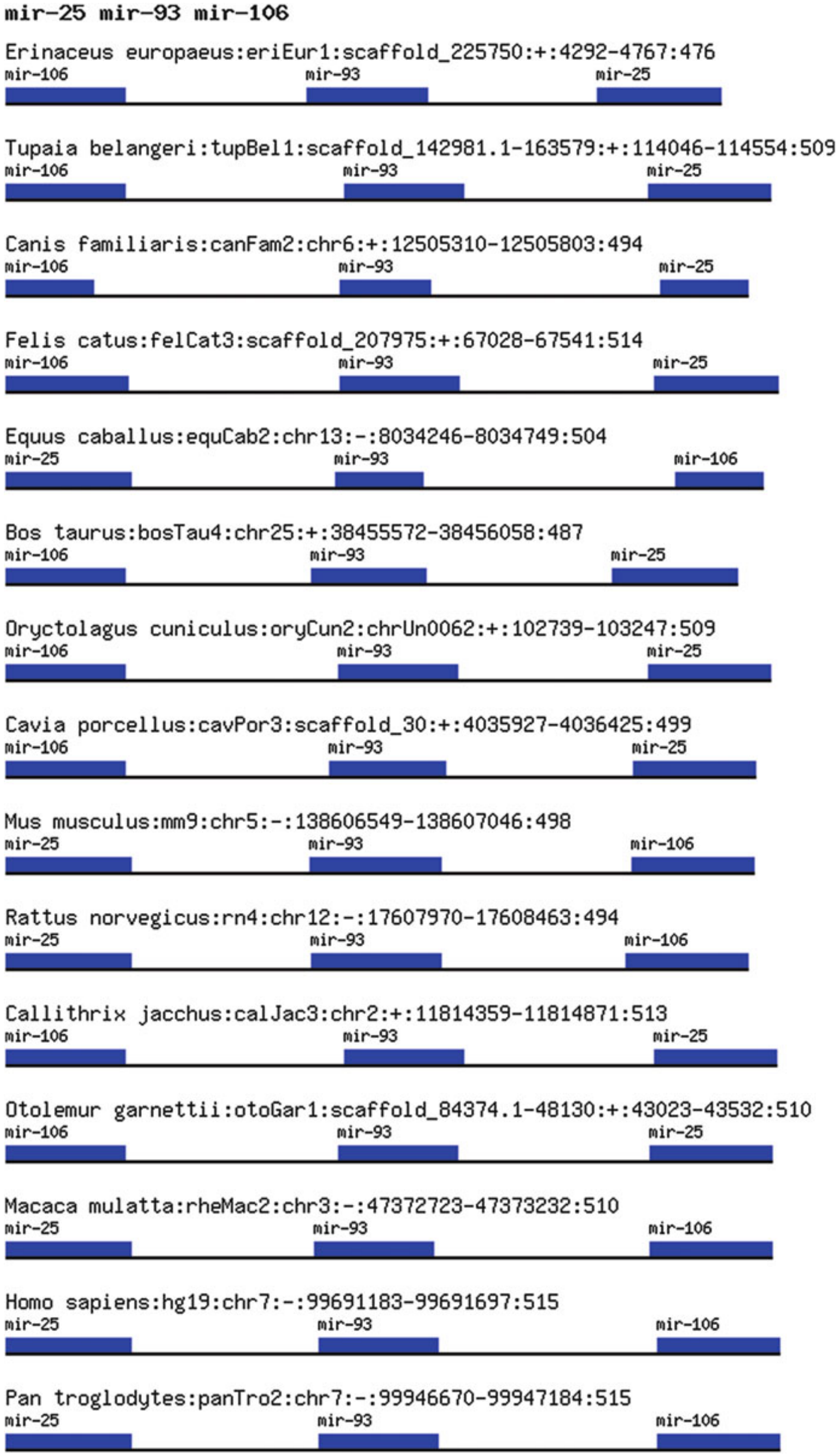

mir-25 mir-93 mir-106
Erinaceus europaeus:eriEur1:scaffold_225750:+:4292-4767:476
mir-106
mir-93
mir-25
Tupaia belangeri:tupBel1:scaffold_142981.1-163579:+:114046-114554:509
mir-106
mir-93
mir-25
Canis familiaris:canFam2:chr6:+:12505310-12505803:494
mir-106
mir-93
mir-25
Felis catus:felCat3:scaffold_207975:+:67028-67541:514
mir-106
mir-93
mir-25
Equus caballus:equCab2:chr13:-:8034246-8034749:504
mir-25
mir-93
mir-106
Bos taurus:bosTau4:chr25:+:38455572-38456058:487
mir-106
mir-93
mir-25
Oryctolagus cuniculus:oryCun2:chrUn0062:+:102739-103247:509
mir-106
mir-93
mir-25
Cavia porcellus:cavPor3:scaffold_30:+:4035927-4036425:499
mir-106
mir-93
mir-25
Mus musculus:mm9:chr5:-:138606549-138607046:498
mir-25
mir-93
mir-106
Rattus norvegicus:rn4:chr12:-:17607970-17608463:494
mir-25
mir-93
mir-106
Callithrix jacchus:calJac3:chr2:+:11814359-11814871:513
mir-106
mir-93
mir-25
Otolemur garnettii:otoGar1:scaffold_84374.1-48130:+:43023-43532:510
mir-106
mir-93
mir-25
Macaca mulatta:rheMac2:chr3:-:47372723-47373232:510
mir-25
mir-93
mir-106
Homo sapiens:hg19:chr7:-:99691183-99691697:515
mir-25
mir-93
mir-106
Pan troglodytes:panTro2:chr7:-:99946670-99947184:515
mir-25
mir-93
mir-106

Fig. 6 The evolutionarily conserved patterns of *mir-25-106*. This figure shows the conservation pattern of the *mir-25-106* (7q22.1) across 15 species according to the proportion of genomic distance

from joining tables by using the SQL language. For our predicted miRNA genes across metazoans, the dumped data from ZooMir (http://insr.ibms.sinica.edu.tw/ZooMir/ZooMir.Candidates_3.tar.bz2) can be easily incorporated into the latest version of miRBase.

4.2 Understanding the Basics of Data Mining and Machine Learning

In recent years with the large-scale and genome-wide data generated by ever-developing molecular biology technique, the huge amount of data have become the major challenge for biologists to manipulate and analyze them using conventional approaches. Increasing evidence suggests that data mining and machine learning approaches can facilitate researchers to efficiently and effectively conquer the massive number of data like in biological research. For instance, in MetaMirClust we introduced a data mining approach to efficiently discover highly conserved sets of miRNA genes upon miRNA clusters. By treating miRNA genes as items, FP-growth algorithm can be utilized to mining the frequent item sets without using candidate generations, of which it can dramatically improve performance in terms of memory space and running time. The algorithm first compresses the input data into a tree-based structure, FP-tree, in which all frequent item sets can be retrieved after easily tracing the entire tree. By iteratively tracing the sub FP-tree based on conditional frequent item sets, the algorithm can efficiently reduce the search costs by avoiding the problem introduced in other approaches to look for short fundamental patterns recursively. Subsequently, the identified frequent item sets using the FP-growth algorithm are equivalent to the frequently co-occurred miRNA genes in terms of clusters. Based on those conservation sets of miRNA genes, we can further reconstruct the hierarchical structure of conservation patterns across metazoans to facilitate the community users to gain more insights into the recruitment process of miRNA genes in clusters in evolution perspective.

4.3 Investigation of Conservation Between miRNA Clusters and Flanking Protein-Coding Genes

To test whether miRNA clusters are co-conserved with their flanking protein-coding genes, we have conducted a downstream analysis, in which the linkage of known protein-coding genes in the vicinity of evolutionarily conserved miRNA clusters between human and mouse were interrogated. We focused only on the nearest adjacent known genes located in the upstream/downstream regions of conserved miRNA clusters upon the same strand between those two species. The genomic information of the protein-coding genes in human (hg19) and mouse (mm9) were downloaded from the UCSC Genome Browser (https://genome.ucsc.edu/). In addition, the liftOver program (http://genome.ucsc.edu/cgi-bin/hgLiftOver) downloaded from UCSC Genome Browser was utilized to find the best mapping of genomic locations between human and mouse if a miRNA cluster occurs in multiple locations. The homologous annotations between known protein-coding genes were identified according to the HomoloGene release 64 from NCBI (http://www.ncbi.nlm.nih.gov/homologene). As a result, our result demonstrated that 24 out of 37 genomic regions were co-conserved according to the evolutionarily conserved miRNA clusters and their corresponding adjacent protein-coding genes. Nine out of thirty-seven genomic regions were partially con-served with either upstream or downstream protein-coding genes. Intriguingly, all six conserved miRNA clusters located in the intronic regions were entirely conserved with their host protein-coding genes. This may suggest that the conservation pattern could be largely extended from miRNA clusters to their adjacent protein-coding genes.

References

1. Ambros V (2004) The functions of animal microRNAs. Nature 431(7006):350–355
2. Bartel DP (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. Cell 116 (2):281–297
3. He L, Hannon GJ (2004) MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet 5(7):522–531
4. Lee Y et al (2002) MicroRNA maturation: stepwise processing and subcellular localization. EMBO J 21(17):4663–4670
5. Lee RC, Feinbaum RL, Ambros V (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 75(5):843–854
6. Reinhart BJ et al (2000) The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. Nature 403(6772):901–906
7. Tsai KW et al (2010) Epigenetic regulation of miR-196b expression in gastric cancer. Genes Chromosomes Cancer 49(11):969–980
8. Tsai KW et al (2011) Aberrant hypermethylation of miR-9 genes in gastric cancer. Epigenetics 6(10):1189–1197
9. Tsai KW et al (2012) Aberrant expression of miR-196a in gastric cancers and correlation with recurrence. Genes Chromosomes Cancer 51(4):394–401
10. Tsai KW et al (2011) Epigenetic regulation of miR-34b and miR-129 expression in gastric cancer. Int J Cancer 129(11):2600–2610
11. Berezikov E (2011) Evolution of microRNA diversity and regulation in animals. Nat Rev Genet 12(12):846–860
12. Griffiths-Jones S (2004) The microRNA registry. Nucleic Acids Res 32(Database issue): D109–D111
13. Li SC et al (2010) Discovery and characterization of medaka miRNA genes by next generation sequencing platform. BMC Genomics 11 (Suppl 4):S8
14. Li SC et al (2010) Identification of homologous microRNAs in 56 animal genomes. Genomics 96(1):1–9
15. Wu HH, Lin WC, Tsai KW (2014) Advances in molecular biomarkers for gastric cancer: miRNAs as emerging novel cancer markers. Expert Rev Mol Med 16:e1
16. Altuvia Y et al (2005) Clustering and conservation patterns of human microRNAs. Nucleic Acids Res 33(8):2697–2706
17. Yu J et al (2006) Human microRNA clusters: genomic organization and expression profile in leukemia cell lines. Biochem Biophys Res Commun 349(1):59–68
18. Sewer A et al (2005) Identification of clustered microRNAs using an ab initio prediction method. BMC Bioinformatics 6:267
19. Hertel J et al (2006) The expansion of the metazoan microRNA repertoire. BMC Genomics 7:25
20. Chan WC et al (2012) MetaMirClust: discovery of miRNA cluster patterns using a data-mining approach. Genomics 100(3):141–148
21. Mathelier A, Carbone A (2013) Large scale chromosomal mapping of human microRNA structural clusters. Nucleic Acids Res 41 (8):4392–4408
22. Tanzer A, Stadler PF (2004) Molecular evolution of a microRNA cluster. J Mol Biol 339 (2):327–335
23. Calin GA, Croce CM (2006) MicroRNA signatures in human cancers. Nat Rev Cancer 6 (11):857–866
24. Laddha SV et al (2013) Genome-wide analysis reveals downregulation of miR-379/miR-656 cluster in human cancers. Biol Direct 8:10
25. Itesako T et al (2014) The microRNA expression signature of bladder cancer by deep sequencing: the functional significance of the miR-195/497 cluster. PLoS One 9(2):e84311

26. Hayashita Y et al (2005) A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. Cancer Res 65(21):9628–9632
27. Yoshino H et al (2013) Tumor-suppressive microRNA-143/145 cluster targets hexokinase-2 in renal cell carcinoma. Cancer Sci 104(12):1567–1574
28. Esquela-Kerscher A, Slack FJ (2006) Oncomirs: microRNAs with a role in cancer. Nat Rev Cancer 6(4):259–269
29. Zhang Y, Zhang R, Su B (2009) Diversity and evolution of MicroRNA gene clusters. Sci China C Life Sci 52(3):261–266
30. Han JW et al (2004) Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Min Knowl Discov 8(1):53–87
31. Chen L, Liu W (2013) Frequent patterns mining in multiple biological sequences. Comput Biol Med 43(10):1444–1452
32. Megraw M et al (2007) miRGen: a database for the study of animal microRNA genomic organization and function. Nucleic Acids Res 35 (Database issue):D149–D155
33. Lai EC et al (2003) Computational identification of Drosophila microRNA genes. Genome Biol 4(7):R42
34. Lagos-Quintana M et al (2003) New microRNAs from mouse and human. RNA 9(2): 175–179
35. Berezikov E et al (2005) Phylogenetic shadowing and computational identification of human microRNA genes. Cell 120(1):21–24
36. Alexiou P et al (2010) miRGen 2.0: a database of microRNA genomic information and regulation. Nucleic Acids Res 38(Database issue): D137–D141
37. Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 42(Database issue):D68–73