國立交通大學

生物資訊及系統生物研究所

博士論文

利用次世代定序技術系統化分析微小非編碼核醣核酸

Systematic analysis of small non-coding RNA using next-generation sequencing technology

研究生　：王威霽

指導教授：黃憲達 教授

中華民國一百年八月

利用次世代定序技術系統化分析微小非編碼核醣核酸

Systematic analysis of small non-coding RNA using next-generation sequencing technology

研究生　：王威齊　　Student : Wei-Chi Wang

指導教授：黃憲達　　Advisor : Hsien-Da Huang

國立交通大學
生物資訊及系統生物研究所
博士論文

A Thesis

Submitted to Institute of Bioinformatics and Systems Biology

College of Biological Science and Technology

National Chiao Tung University

for the doctoral degree

in bioinformatics and systems biology

August 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年八月

利用次世代定序技術系統化分析微小非編碼核醣核酸

學生:王威霽　　　　　　　　　　指導教授:黃憲達 博士

國立交通大學 生物資訊及系統生物研究所博士班

摘要

次世代定序技術(NGS)是一個新的定序技術，它是一個能夠快速、大量以及穩定偵測和分析微小非編碼核醣核酸(small non-coding RNA)的表現量的方法。微小非編碼核醣核酸在生物體內主要影響到基因沉默(gene silencing)、DNA 甲基化、組蛋白修飾。為了研究微小非編碼核醣核酸的功能，微陣列(microarray)被廣泛的使用。隨著次世代定序的發展，近年來越來越多生物學家利用次世代定序技術來研究微小非編碼核醣核酸的功能。次世代定序比微陣列在分析微小非編碼核醣核酸上更為準確且能發現未知的微小非編碼核醣核酸。次世代定序技術提供生物學家更好的資料品質，但是到目前為止，如何廣泛且完整的分析微小非編碼核醣核酸還未被提出。這個研究的主要目標就是針對次世代定序技術在微小非編碼核醣核酸建立廣泛且完整的分析流程，希望這個系統化的分析流程能夠幫助生物學家快速的分析資以及獲得更多有用的結果。

Systematic analysis of small non-coding RNA using next-generation sequencing technology

Student : Wei-Chi Wang **Advisor : Hsien-Da Huang**

Abstract

Next-generation sequencing (NGS) technology which is a novel sequencing technology offers high-throughput and robust approaches for monitoring and analyzing the expression of small non-coding RNAs. Small non-coding RNAs contain microRNAs (miRNAs) and short interfering RNAs (siRNAs) which play an important role in gene silencing, DNA methylation and heterochromatic histone modifications in animals and plants by targeting mRNAs. Unlike capillary-based sequencing, next-generation sequencing technology produces million of sequences (35~1000 nt) at a time. This offers the quantitative analysis in the small non-coding RNA. Moreover, comparing oligonucleotide microarray, it can be used to not only profile the expression levels of known miRNAs but also discovery novel miRNAs. In the recent years, there are more and more studies applying this sequencing technology to deciphering the function of small non-coding RNAs. However, there are no complete and comprehensive analysis flows for these sequencing data.

This work aims at developing the systematic and comprehensive analysis pipeline in small non-coding RNA including animals and plants for next-generation sequencing data. This pipeline can help biologists to easily apply their own NGS data to finding the biological significant small non-coding RNAs, pathways and regulation networks between small non-coding RNAs and their target genes.

致謝

關於這個研究首先要感謝中研院植物暨生物學研究所林納生博士於 2008 年提供第一份 NGS data 開啟我投入這個研究。在當時台灣很少有 NGS data 或是相關研究，透過和實驗室林豐茂學長合作研究了解如何處理 NGS data，接著發表第一篇論文。透過處理 NGS data 以及我們 lab 研究 miRNA 的專長，陸續和 UCR 的史允中以及 Hailing Jin 教授合作分析 NGS small non-coding RNA data 在動物和植物上，都有不錯的成果。

關於 miRNA 的一些相關知識要特別感謝勝達常常給我在 miRNA target prediction 上的建議，Alan 在 miRNA TSS 方面的 support，熙淵學長在 promoter analysis 的建議和指導。還有博凱常常在我需要一些資料時給予我協助。除此之外，還要感謝實驗室的學弟妹在一些瑣碎的事上幫忙。

當然最感謝的是我的指導教授黃憲達博士對我從碩士班到博士班的包容以及指導，最後感謝我的父母以及哥哥姐姐對我的支持和幫忙，讓我能安心完成學業。

Table of contents

List of Figures

List of Tables

1. Introduction

Motivation and Specific Aims

Non-coding small RNAs like microRNAs (miRNAs) and short-interfering RNAs (siRNAs) play an important role in gene regulation by mRNA degradation, translation repression and transcriptional gene silencing in animals and plants. To investigate the function of small RNAs, the method for high-throughput monitoring their expression level is needed. The oligonucleotide microarray is one of common used method for detecting the expression of small RNAs. But this method can only detect known small RNA sequences by designing corresponding probe sets. To overcome this problem, next-generation sequencing (NGS) technology which can produce millons of sequencing reads at a time is applied for profiling the expression level of small RNAs. NGS has higher sensitivity and specificity than olignucleotide microarray. Moreover, NGS can identify novel non-coding small RNAs such as novel miRNAs. There are over 60 published studies about non-coding small RNA analysis by next-generation sequencing technology in animals, plants, bacterial and virus (Table 1). The amount of published studies per year also increases from the date (2004) which NGS is available (Figure 1).

To handle large amount of small RNA sequencing data sets, various standalone packages, web-based tools and algorithms are designed for different purposes. For example, using traditional alignment tools like BLAST [1] and blat [2] to map sequencing data to genomic sequences is not efficient. So, several tools such as SeqMap [3], ZOOM [4], Maq [5], Bowtie [6] and SOAP [7] are developed for analyzing numerous short sequences and map millions of short sequences to genomic sequences. miRDeep [8] is designed for identify novel miRNAs from small RNA NGS data. For automatically profiling the expression of miRNAs, miRExpress [9] and miRNAkey [10], standalone packages, are developed. Web-based tools such as miRanalyzer [11], SeqBuster [12], DSAP [13] and mirTools [14] are also built. However, there are no any systematic and comprehensive analysis flows for annotating small RNA sequencing data completely. In order to solve this problem and provide more detailed annotation for sequencing data, the analysis pipeline of non-coding small RNA data by NGS in different organisms such as animals and plants is certainly need to be constructed.

This work aims at developing the systematic and comprehensive analysis pipeline in small non-coding RNA including animals and plants for next-generation sequencing data. This pipeline contains sequencing raw data preprocessing, the expression level of small RNAs profiling, novel miRNAs discovery, the target genes of small RNAs identification and functional analysis of target genes by combining multiple related data such as Gene Ontology (GO), pathway information from KEGG [15] and gene expression data from cDNA microarray. The method for each part is designed for optimizing the running time and the results like generating more correct expression profiles and reducing the false positive rate of target gene prediction. In addition, easily understood analysis results of each part are also important for researchers. Therefore, the analysis result of this pipeline is designed for abundance of biological significance, easily read and simple formats.

Table 1. Non-coding small RNA studies using next-generation sequencing technology

Species	Number of published studies	References
Animals		
Homo sapiens	19	[16-34]
Mus musculus	2	[35-36]
Others	13	[37-49]
Plants		
Arabidopsis	8	[50-57]
Oryza sativa	3	[58-60]
Others	13	[61-73]
Bacterial	4	[74-77]
Virus	2	[78-79]

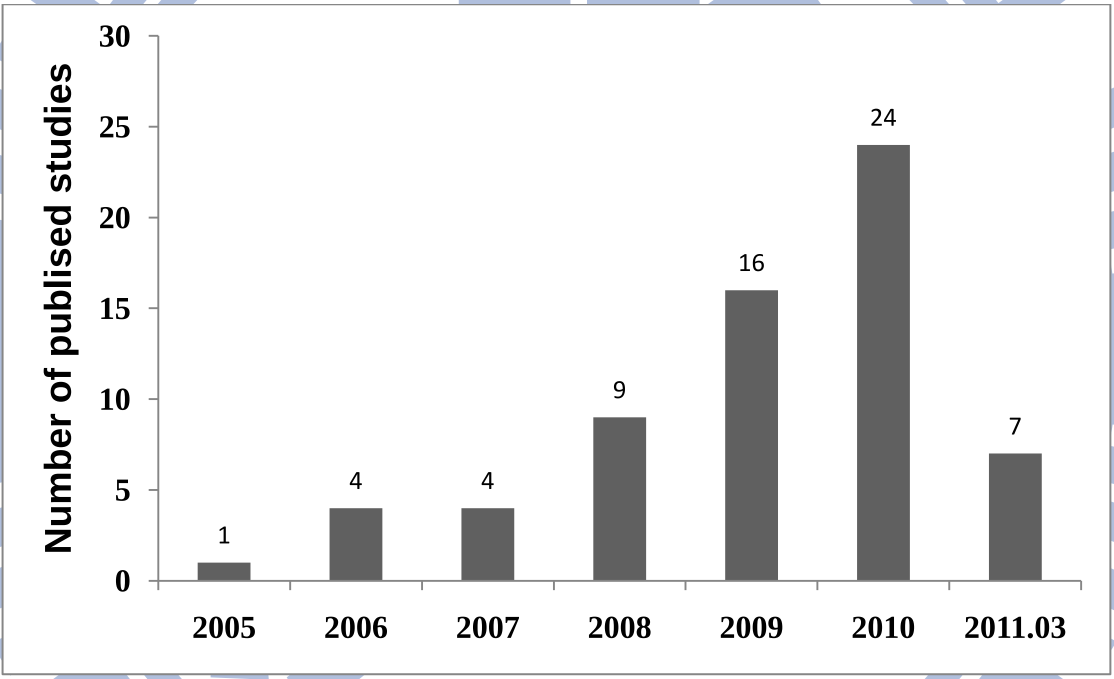


Figure 1. The distribution of non-coding small RNA studies by NGS. Data collection until 2011.03

2. Background

2.1 Next-generation sequencing technology

Next-generation sequencing (deep sequencing) technology is a new sequencing approach which produces millions of sequencing reads at a time. The general workflow of next-generation sequencing is preparing samples for sequencing by ligating specific adaptor oligos to 5’ and 3’ ends of DNA fragments. Then, the samples are subjected to the next-generation sequencers. The length of sequencing reads is 25~1000 nt according to different sequencing platforms. The quantities of sequencing reads are also different in each sequencer. The difference between next-generation sequencing (NGS) and Sanger sequencing are the number of produced reads at a time (NGS: 100 MB-30 GB, Sanger: 0.1 MB), the length of reads (NGS: 25-1000 nt, Sanger: 700-900 nt) and the cost per GB (NGS: $2K~84K, Sanger: >$2500K). Five next-generation sequencers can be chosen now. They are Roch454 FLX sequencer, Illumia Genome Analyzer, Applied Biosystems SOLiD sequencer (ABI SOLiD), HeliScope Single Molecule sequencer (Helicos tSMS) [80-82] and Pacific Biosciences [83].

Roche 454

Roche 454 is the first next-generation DNA sequencer which is commercially introduced in 2004. Roche 454 sequencer produces sequencing reads based on the principle of prosequencing. Figure S1 demonstrates the workflow of Roche 454 sequencer. The sample is constructed by ligating 454-specific adaptors to DNA fragments. Then, ligated DNA fragments are amplified by bead-based emulsion polymerase chain reaction (em-PCR). After em-RCR, amplified beads are loaded into the picotiter plate (PTP). PTP is the solid surface which contains the single wells for packing beads and enzyme beads. In the PTP, all amplified beads are sequenced by pyrosequencing reactions. Nucleotides are flowed sequentially and sequenced by detecting the light which is generated through the release of pyrophosphate. Roche 454 sequencer can produce 700 Mb of sequences per run in 23 hours. The length of sequences is 700-1000 nt [80-82].

Illumia Genome Analyzer

Illumina Genome Analyzer which is developed based on the concept sequencing by synthesis (SBS) is available in 2006. Figure S2 is the workflow of Illumina Genome Analyzer. Before loading the sample into the flow cell, it needs to be done fragmentation and 5’ and 3’ adaptor ligation. The ligated DNA fragments are amplified by bridge amplification (an isothermal process that amplifies each fragment into a cluster). To sequence each cluster in the same direction, one strand of amplified clusters is selectively removed. Then, the flow cell is transferred to the Genome Analyzer. Each single-strand cluster is sequenced by SBS reactions (imaging, removing the fluorescent group and deblocking the 3’ end for next cycle). In each chemistry cycle, only a single base is identified. So, the length of sequencing read is determined by the number of cycles of nucleotide incorporation, image and cleavage. Illumina Genome Analyzer can produce 95 GB of 100-150 nt sequencing reads per run in 2 days [80-82].

ABI SOLiD

Applied Biosystems SOLiD sequencer (ABI SOLiD) is commercial release in October 2007. Like Roche 454, it uses em-PCR to amplify fragment DNA into beads. ABI SOLiD uses a unique sequencing process catalyzed by DNA ligase (Figure S3a). The ligation-based sequencing process begins with annealing a universal primer which is perfect complementary to the 5' end adaptor. Then, 16 random 8-mer probes are added. The first and second 3' end of the probes are labelled using one of four fluorescent dyes and complementary to the template sequences. After ligation, the fifth dinucleotide is imaged. Then, the 6-8 nucleotides of the probes are removed and adding the random probe set. Every dinucleotide profile is imaged after several rounds of ligation and the sequencing reads are identified (Figure S3b). This ligation-based approach is called as 2 base encoding. SOLiD sequencer can produce 10-20 GB of sequencing reads (25-75 nt) per run in 2-4.5 days [80-82].

Helicos' tSMS

Helicos' tSMS sequencing platform is available in 2008. It is the first next-next generation (3rd generation) sequencing platform. The technology of tSMS is different with Roche 454, Illumia and SOLiD. tSMS do not amplify the templates before sequencing. Library preparation is also different with other three sequencing technologies (only adding a poly-A tail and labeling the fluorescent). Then, tailed templates are detected according to the fluorescent label through hybridiinge to poly-T oligonucleotides on the flow-cell surface. The sequencing flow of tSMS is called as terminator SBS. The terminator nucleotides are based on steric hindrance to deter the incorporation of more than one nucleotide per cycle. Fluorescent is removed and the next fluorescent nucleotide is added singly per cycle after identifying incorporate nucleotides. tSMS currently produces 21-28 GB of sequencing reads (25-50 nt) per run in 8-9 days [81].

Pacific Biosciences

Pacific Biosciences is the second 3rd generation sequencing platform (announced in 2010). Like Helicos' tSMS, it does not need to do amplification before sequencing. The sequencing method of Pacific Biosciences is called as single-molecule real-time (SMRT) sequencing. It directly observed DNA synthesis on single DNA molecules in real time by using zero-mode waveguide (ZMW) technology. The first commercial SMRT array contained ~75000 ZMWs. Each ZMW contains a DNA polymerase loaded with DNA samples. The sequencing length of SMRT is >1000 nt (maximum length is more than 10000 nt). Moreover, the time per run is only several hours [83].

Comparison of next-generation sequencing technologies

Table 2 lists the performance comparison of current different next-generation sequencing platforms. Among these five sequencing platforms, tSMS and Pacific Biosciences are 3rd generation sequencing platform which do not do amplification. Therefore, the sequencing error rate is lower than other three platforms (the error rate of Pacific Biosciences could not be obtained from official site but it should be very low according to the sequencing method). This is because some templates which are sequenced do not incorporate a nucleotide at the corresponding cycle during the amplification processes. For Pacific Biosciences, the length of sequencing read is more than 1000 nt. It is more suitable than other sequencing platforms in *de novo* assembly and SNP identification.

Table 2. The performance comparison of next-generation sequencing technologies

	Roche 454 (FLX-Titanium)	Illumia Genome Analyzer (IIx)	ABI SOLiD	Helicos tSMS	Pacific Biosciences
Method of amplification	Bead-based/ emulsion PCR	Bridge amplification	Bead-based/ emulsion PCR	N/A	N/A
Sequencing chemistry	Pyrosequencing	Polymerase-based sequencing-by-synthesis	Ligation-based sequencing	Virtual terminator SBS	Single-molecule real-time sequencing
Read length	700-1000	100-150	35-75	25-55	>1000
Run time	23h	2 days for 36-cycle single-end run, 4 days for 36-cycle paired-end run	6-7 days for fragment libraries, 8 days for 2×25 base paired-end libraries	8-9 days	< 1 day
Throughput/run (Gb)	0.7	95	10-20	21-35	Unknown
Error rate of sequencing	0.005%	0.001%	0.01%	0.005%	Unknown

2.2 The biogenesis of microRNAs in animals

MicroRNAs (miRNAs), small non-coding RNAs of 19~25 nt sequences, play important roles in gene regulation in animals and plants. Generally, miRNAs hybridize to the 3'-untranslated region (3'-UTR) of mRNA to down-regulate gene expression or to induce the cleavage of mRNA and can fully hybridize to the transcripts of target gene [84]. Previous studies have suggested that miRNAs are strongly associated with various cancers [85] and various crucial cell processes such as apoptosis, differentiation and development.

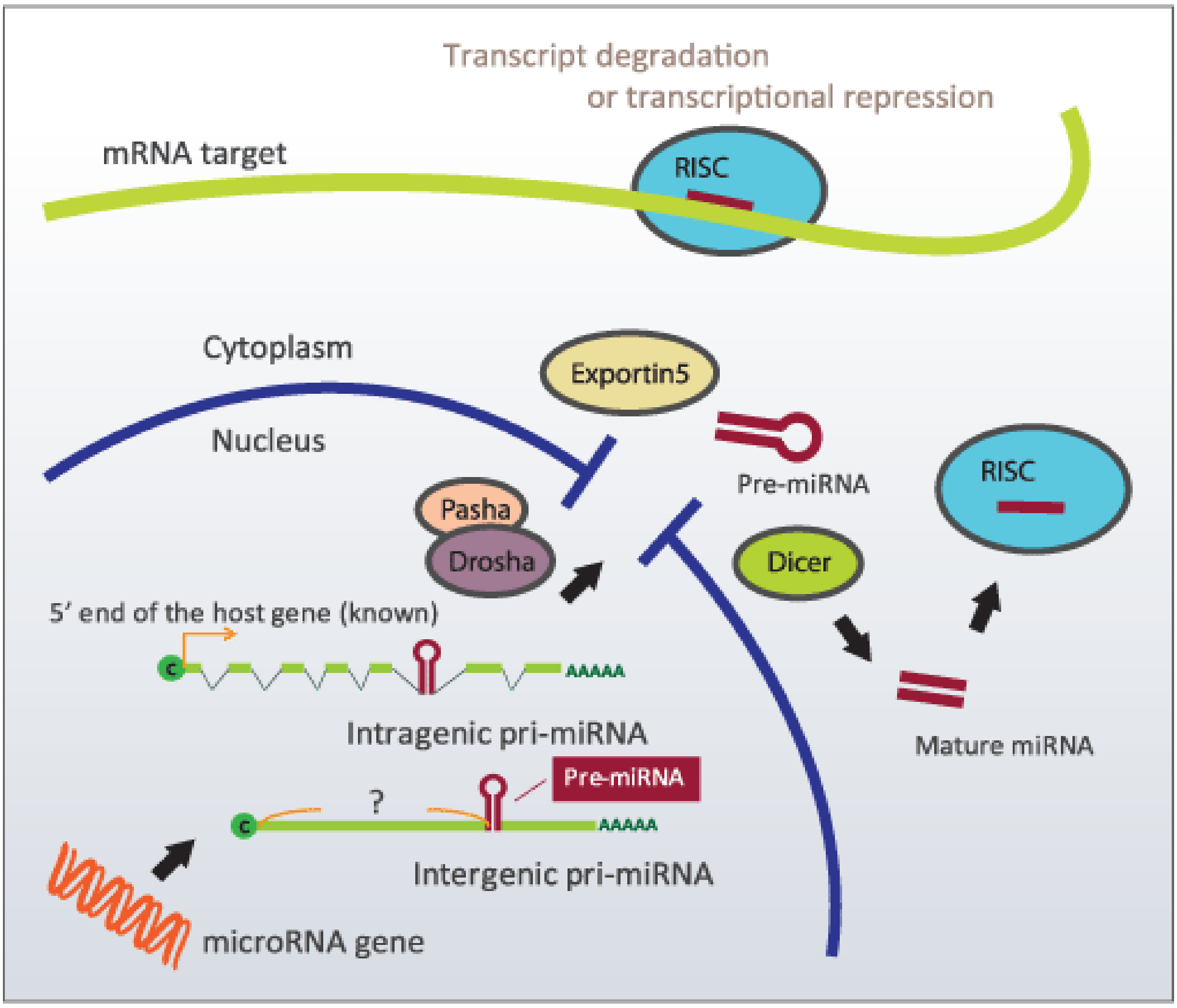


Figure 2. The biogenesis of miRNAs

MicroRNA (miRNA) genes are generally transcribed by RNA Polymerase II (Pol II) in the nucleus to form large pri-miRNA transcripts, which are with 5’ 7-methyl guanosine cap and 3’ poly-A tail. The pri-mRNAs can be classified into two types according to their genomic locations (Figure 2). One is intragenic pri-miRNA which locates in the known gene. Another is intergenic pri-miRNA which locates in the region which does not be annotated as the gene. 60-100 nucleotide pre-miRNA precursors are produced through the transcription process of pri-miRNA transcripts. This transcription is processed by the RNase III enzyme Drosha and its co-factor DGCR8 [86] (also known as Pasha). Drosha functions as the catalytic subunit, while DGCR8 recognizes the RNA substrate in this process. Then, the pre-miRNA is exported to the cytoplasm by exportin-5 and turned into the miRNA duplex by RNase III enzyme Dicer [87]. One strand of the miRNA duplex is incorporated into the protein complex known as RNA-induced silencing complex (RISC) with Argonaute proteins [88]. Previous study shows that nuclear RISC, consisting only of Ago2 and mature miRNA, is loaded in the cytoplasm and imported into the nucleus [89]. Diederichs and Haber indicate that Ago2 can serve as Dicer enzyme to cleave pre-miRNA [90]. The mature miRNA then binds to complementary sites in the 3' untranslated regions (3'-UTRs) of mRNA to negatively regulate gene expression depend on the degree of complementarity between the miRNA and its target. miRNAs that bind to mRNA targets with imperfect complementarity block the expression of target gene at the level of protein translation. Complementary sites for miRNAs by this mechanism are generally found in the 3’UTR of the target mRNA genes. miRNAs that bind to their mRNA targets induce target-mRNA cleavage. The regions among miRNA-binding sites called as seed regions (Watson–Crick consecutive base pairing between mRNAs and the miRNA at position 2-8 from its 5' end) [91] located in 3'-UTRs of mRNAs are important to translational repression and mRNA degradation [92].

Figure 3 shows the function of miRNAs in different ways. Plant miRNAs differ from animal miRNAs in that many plant miRNAs have perfect homology to their target mRNAs, and they act through the RNAi pathway to cause mRNA degradation [93]. However, some plants with imperfect complementarity to their target sites and act a function similar to animal miRNAs. Plant miRNAs are also known to target chromatin modifications, such as histone methylation and DNA methylation [94].

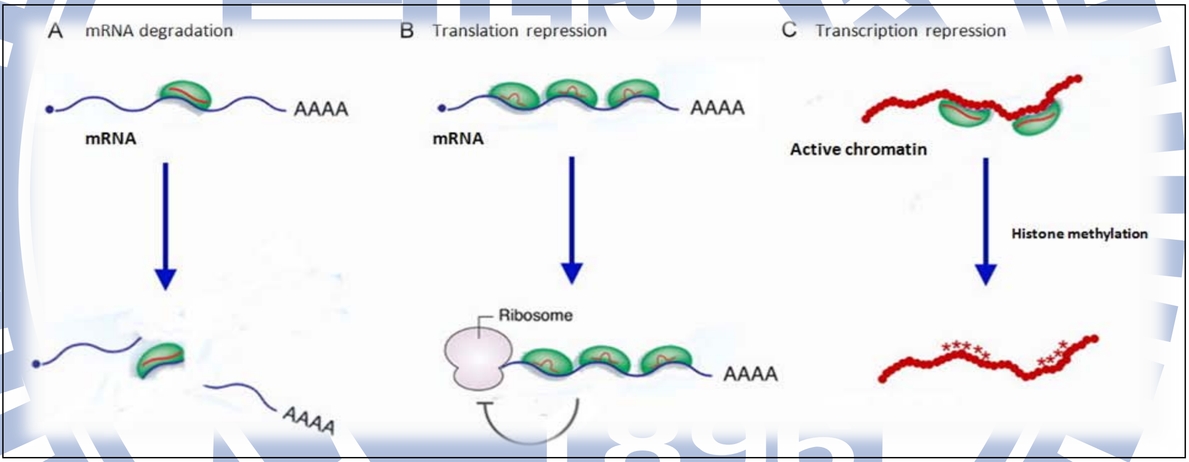


Figure 3. miRNA function: (A) mRNA degradation, common in plants, (B) Translation repression, common in animals, (C) Transcription repression by histone or DNA methylation, common in yeasts, plants, and possibly animals

2.3 MicroRNAs and siRNAs in plants

In plants, small RNAs (19-27 nt) can be classified into two classes, microRNAs (miRNAs) and short interfering RNAs (siRNAs), play an important role in gene regulation. siRNAs contain trans-acting siRNAs (tasiRNAs), natural antisense transcript siRNAs (nat-siRNAs) and heterochromatic siRNAs (hc-siRNAs). Figure S4 demonstrates that the biogenesis and relationship of miRNAs and each type of sRNA in plants.

The biogenesis of miRNAs in plants

The biogenesis of miRNAs in plants has some difference with animals. In plants, the length of stem-loops generally is longer than that of animal stem-loops. DCL1 (DICER-LIKE1) has the functions of Drosha and Dicer. The miRNA/miRNA* duplex is produced through DCL1 interacting with HYL1 (HYPONASTIC LEVAES1) to cleave the miRNA precursor in nucleus [95-98]. The mature miRNA is methylated by HEN1 (HEN1 is also in nucleus) [94]. Methylation protects small RNAs from degradation and polyuridylation. Exporting miRNA duplexs or mature miRNAs to the cytoplasm is completed by HASTY [99] (the function is similar with Exportin-5) (Figure S5).

Like animals, plant miRNAs require Argonaute protein for forming RNA-induced silencing complex (RISC) [100]. However, the degree of complementarity between miRNAs and their targets is different in plants. Unlike animals, nearly perfect complementarity between miRNAs and their targets is required in plants [100-103]. Most plants miRNAs only have four or less mismatches with their targets. These mismatches usually locate in the 3’ region of miRNAs [104]. The degree of complemnetarity in the region of miRNA/mRNA duplex at position 3-11 affects the efficiency of cleavage in mRNAs [105]. The affection of miRNAs and their targets is also different between animals and plants. In animals, most miRNAs guide translational repression of their target genes. In plants, most miRNAs lead to mRNA cleavage by targeting the coding region of mRNAs. The previous studies demonstrated that about two-thirds of plant miRNAs regulate the expression of transcription factors during plant development. In additional to their great affection in development [106-107], they also play important roles in plant responses to biotic and abiotic stresses and nutrient homeostasis [106, 108-112].

The biogenesis of tasiRNAs

Trans-acting siRNAs (tasiRNAs) which has similar function to miRNAs down-regulate the expression of genes are 21 nt regulatory siRNAs. TasiRNAs are generated from specific miRNAs cutting TAS primary RNAs process. TAS genes transcribe long primary RNAs which do not generate protein products. Its' function is serving as the precursors of the tasiRNA. In processing tasiRNAs, miRNAs guide the cleavage of tasiRNA primary transcripts which are converted into the structure of dsRNA by SGS3 and RDR6 binding to one of the two single-stranded TAS cleavage sequences. Then, DCL4 (DICER-LIKE4) with DRB4 cuts the dsRNAs and produces 21 nt tasiRNA-mRNA duplexes (Figure S6) [113-115]. Like miRNAs, tasiRNAs are methylated by HEN1 for avoiding degradation and polyuridylation. Different tasiRNA families regulate the gene expression by forming RISC with different AGO family protein (Figure S4). AGO1 involves with TAS2 tasiRNAs and AGO7 involves with TAS3 tasiRNAs. The high degree of complementarity between tasiRNAs and mRNA is needed to guide mRNA cleavage. The different members of same gene family are targeted by either miRNAs or tasiRNAs. For example, miR161 or TAS2 tasiRNA target the members of PPR family and miR160, miRN167 or TAS3 tasiRNA target the members of ARF family, respectively. According to previous studies report, tasiRNAs are found only in plants [116].

The biogenesis of nat-siRNAs

Natural antisense transcript siRNAs (nat-siRNAs), 24 or 21 nt siRNAs, are generated from natural antisense transcripts (NATs). NATs are formed by two coding or non-coding RNAs that have complementary regions. There are two classes of NATs, cis-NATs and trans-NATs [50, 117-125]. The transcripts in the same genomic locus but in different strands form cis-NATs. For example, the genomic locus of SRO5 is at chr5: 25097998-2509997 [+]. The genomic locus of P5CDH is at chr5: 25099003-25103298 [-]. The overlapping region is at chr5: 25099003-2509997. The cis-NATs can be categorized into three groups, convergent (overlapping in the 3' ends of two transcripts), divergent (overlapping in the 5' end of two transcripts) and enclosed (one transcript can completely overlap another transcript) according to the conditions of overlapping. Trans-NATs are formed by the overlapping regions of two transcripts from different genomic locus.

The 24 nt nat-siRNAs are derived from the complementary region by the interaction of DCL2 (DICER-LIKE2), NRPD1a, RDR6 and SGS3. These 24 nt nat-siRNAs guides the cleavage of the constitutive transcript and establishes a phase for the sequential production of 21 nt nat-siRNAs by DCL1 (Figure 7). Unlike tasiRNAs, the function of nat-siRNAs is not for targeting other mRNAs. It can lead to post-transcriptional gene silencing by hybridizing to the cis-strand of mRNA. The member of AGO family involves in this mechanism does not be understood clearly. The well known example is SRO5 and P5CDH in *Arabidopsis*.

SRO5 and P5CDH have the function of regulating salt tolerance [126]. SRO5 and P5CDH have the overlapping region in their 3'end of transcripts. P5CDH is expressed constitutively. When SRO5 is induced in response to salt stress, the 24 nt nat-siRNAs are derived from the overlapping region of these two transcripts by DCL2, RDR6 and SGS3 interaction. These siRNAs direct the cleavage of P5CDH transcripts (Figure S7). Meanwhile, the dsRNAs are formed by RDR6, SGS3 and SDE4. 21 nt nat-siRNAs are produced from the cleavage of dsRNAs by DCL1. Then, the mRNA of P5CDH are degraded through these 21 nt nat-siRNAs targeting P5CDH.

The biogenesis of trans- nat-siRNAs does not be understood clearly. However, previous studies suggests that trans- NAT-siRNAs are involved in alternative splicing, post-transcriptional gene silencing [127-129]. There are some studies using the computational methods to screen whole *Arabidopsis* transcripts to find trans-NATs.

The biogenesis of hc-siRNAs

Heterochromatic siRNAs (hc-siRNAs), 24 nt siRNAs, are produced from transposable or repetitive elements [130-132]. Single strand non-coding transcripts are transcribed from heterochromatic locs by Pol IV and CLASSY1 (Figure S8). RDR2 involves in the formation of dsRNAs through using transcripts as templates. Then, the dsRNAs are processed into 24 nt siRNAs by DCL3. Methylation by HEN1 joins the process for protecting degradation and polyuridylation. The siRNAs are loaded into AGO4-RISC complex. DNA methylation and heterochomatic histone modifications are affected by PoI V and AGO4 interaction and the interaction in DRD1 (potential chromatin remodeling protein), DMS3 (structural maintenance of chromosomes hinge-domian protien) and Pol V [133-137]. In *Arabidopsis*, the siRNAs derived from a transposable element inserting in the intron of FLC (FLOWERING LOCUS C) genes lead to the reduction of FLC and early flower. FWA gene is silenced through the siRNAs, generated from two tandem repeats in the promoter region of FWA, triggering DNA methylation. The siRNAs derived from the tandem repeats in the promoter region of SDC, a gene ecncoding the F-box protein, silenced SDC by triggering DNA methylation.

2.4 Regulation role of small RNA

The regulation role of miRNA in animals

The miRNAs regulate the gene expression by targeting the 3'-UTR of mRNAs, resulting in degradation of mRNAs and repression of translation. miRNAs are reported strongly associated with various pathways such as cancer pathway and cholesterol metabolic pathway.

The well-known example in cancer pathway is that miRNA-34 (miR-34) family, miR-34 a, b and c, is down-regulated in various cancers. Figure S9 demonstrates the regulation role of miR-34 family in cancer pathway [138]. p53 is the key factor which regulate the abundance of miR-34 in cancer pathway. miR-34 family targets the set of mRNAs which support tumor formation such as E2F3, Bcl2, Notch, HMGA2, CDK4, CDK6 and Cyclin E2. E2F3 is an important transcription factor which initiated the production of proteins that affect cell-cycle checkpoint, DNA repair and replication. miR-34 family inhibits cell proliferation and activates cell death pathways by down-regulating the expression of E2F3. Bcl2 which usually over-expressed in cancer cells is the important gene in tumor formation. It protects tumor cells processed into apoptosis pathway. Reduction of Bcl2 expression level by miR-34 family targeting induces the apoptosis of cancer cells. The other miR-34 target genes such as Notch, HMGA2, CDK4, CDK6 and Cyclin E2 are involved self-renewal and survival of cancer stem cells. Therefore, the miR-34 family can inhibit tumor formation. According to the function of these miR-34 target genes, miR-34 family is a tumor suppressor in cancer pathway.

Recent studies reported miRNAs involve in cholesterol metabolic pathway. miR-33 a and b are the intragenic miRNAs [139-141]. Their host genes are SREBP (sterol regulatory element-binding protein) family. SREBP family is the transcription factor which is the key protein in cholesterol biosynthesis. miR-33 is co-expressed with its host gene (SREBP) and targets ABCA1 (adenosine triphosphate-binding cassette transporter A1). ABCA1 is the important regulator in HDL (high-density lipoprotein) synthesis and reversing cholesterol transport. Figure S10 demonstrates the regulation role of miR-33 in cholesterol metabolic pathway. Cholesterol accumulation induces the expression of ABCA1 and inhibits the expression of SREBP. ABCA1 medicates cholesterol efflux to apolipoprotein A1 (apoA-1) and HDL. The expression of ABCA1 is down-regulated by miR-33. Therefore, the role of miR-33 in cholesterol metabolic pathway is regulating the expression level of HDL with SREBP (the host gene of miR-33) by targeting ABCA1.

The regulation role of small RNAs in plants

In plants, many genes silencing are through small RNA such as miRNAs and siRNAs. miRNAs regulate the expression level of genes by targeting the coding region of the mRNA. siRNAs regulate the gene expression through multiple ways such as directly targeting the coding region of mRNA or transcriptional gene silencing by triggering DNA methylation and histone modifications. Table S1 lists the reported miRNAs which are involved in root development in *Arabidopsis* and Rice [142].

The miRNAs involve in auxin signing pathway are miR160, miR164, miR167, miR390 and miR393. miR160 targets ARF10, ARF16 and ARF17. The function of ARF10 and ARF16 is for root cap development. Overexpressing miR160 causes serious root cap defect in rice (the targeting relationships do not be validated yet). The function of ARF17 is reducing primary root length and decreasing root branching in *Arabidopsis*. miR164 is the key regulator in lateral root development. miR164 targets NAC1 which transduces auxin signal for the emergence of lateral root. So, miR164 down-regulates auxin signal in lateral root development. miR167 targets ARF6 and ARF8 which are positive regulator of shoot-borne root emergence. miR390 regulate the expression of ARF2, ARF3 and ARF4 which involve in auxin signaling pathway by indirect targeting. miR390 target TAS3 which is the non-coding transcript and produce tasiRNAs. These tasiRNAs direct target ARF2, ARF3 and ARF4. miR390 regulates the auxin signal concentration and lateral root development by TAS3-ARF2/ARF3/ARF4 pathway. miR393 targets TIR1 (transport inhibitor response 1), AFB2 (auxin signaling F-Box proteins 2) and AFB3. So, miR393 regulates auxin signaling. Moreover, NAC1, the target gene of miR164, is the downstream of TIR1. NAC1 transduce auxin signal to promote lateral root development. Therefore, miR164 and miR393 may participate together in lateral root development.

Vegetative and reproductive processes of plant root require mineral elements in soil such as macronutrients, micronutrients and heavy metals. Several miRNA families involve in nutrition response or metabolism which affect plant growth. The first miRNA family is miR395 which is the key regulatore in sulphate metabolism by targeting SULRT2;1 (low-affinity sulphate transporter), APS1 and APS2 (ATP sulphurylases). The second miRNA family is miR398 which affects copperand zinc homeostasis by targeting CSD1 and CSD2. CSD1 and CSD2 are copper/zinc superoxide dismutase genes. Heavy metals such as copper and zinc are required for root growth. Therefore, miRNA398 has great influence in plant root development and growth behavior of plants. The final miRNA family is miR399. miR399 involves in phosphate starvation response in *Arabidopsis*. The concentration of phosphate is one of important factor in root system architecture and growth habit. miR399 regulates phosphate homeostasis by targeting PHO2. External abiotic stress such as drought affects plant root growth. miR169 is induced by drought stress in Rice. But the target genes of miR169 are not identified.

miRNAs in plants also involve in other pathway excluding root development [133]. For example, miR156 targets over 50% members of SPL family. SPL 3, 4 and 5 promote the change of vegetative phase and floral transition. The length of plastochron is regulated by SPL9 and SPL15. miR172 targets AP2 (APETALA2) which regulates floral transitions. miR159 targets MYB33, MYB65 and MYB101. These MYB genes induce Gibberellin-responsive (GA) genes in the aleurone layer. Floral pattering and leaf morhpgenesis are regulated through miR164 targeting CUC1 (CUPSHAPED COTYLDON1) and CUC2 (CUPSHAPED COTYLDON2). miR165/166 targets RHB and REV which promote abaxial identity of lateral organs in both *Arabidopsis* and maize. miR319 targets several TCP genes (a large family of proteins containing “TCP domain”). TCP genes regulate cell divisons during leaf morphogenesis.

2.5 Application of NGS in small RNA

Small RNAs play an important role in regulating the expression level of gene. In the last few years, high-throughput and robust approaches for monitoring the expression of miRNAs have been used to understand how miRNAs are differentially expressed under various conditions. The oligonucleotide microarray is one method for detecting miRNA expression. This approach involves the design of probes based on known miRNAs that are collected in miRBase for miRNA expression profiling studies. However, the approach is restricted to detecting the expression of known miRNAs. Next-generation sequencing technology is an inexpensive and high-throughput sequencing method. This new method is a promising tool with high sensitivity and specificity and can be used to measure the abundance of small-RNA sequences in a sample. Hence, the expression profiling of miRNAs can involve use of sequencing rather than an oligonucleotide array. Additionally, this method can be adopted to discover novel miRNAs. There are over 60 published researches which applied NGS to their small RNA studies in animals, plants and bacterial (Table 1). In the latest version of miRBase [143], a database collecting 17341 mature miRNAs, there are 8117 mature miRNAs which are identified by NGS. NGS is a powerful method for profiling miRNAs expression and discovery novel mRNAs (near 50% miRNAs are found by NGS).

3. Related works

3.1 The repositories of miRNAs

miRBase

miRBase (http://www.mirbase.org/) is the first database which collects published miRNA sequences [143]. It contains 15172 miRNA precursors which express 17341 mature miRNAs in 142 species (miRBase 16.0). miRBase offers the comprehensive information for each miRNA such as the chromosome location, the predicted target genes and the expression level detected by next-generation sequencing technology. Users can also check their own sequences with whole miRNAs in miRBase. Moreover, miRBase offers users to download all data in the database for user-defined analysis by themselves.

3.2 miRNA target site prediction tools

TargetScan/TargetScans

TargetScan [144] is the first developed miRNA target site prediction tool based on the concept of seed matches. The definition of seed matches is that seven continuous perfect complementarity in the miRNA /mRNA duplex in the 2-8 positions of the miRNA. A:U, C:G and G:U pairs are considered as complementarity. TargetScan scans the 3'UTR of mRNAs. For each predicted target site, the free energy of the duplex is assigned. TargetScans is the improvement version of TargeScan. TargetScans scans a set of orthologous 3'UTR of mRNA in a group of organisms. In TargetScans, seed matches are classified into 3 main types. The first is 6-nt seed match (6mer). The second is 7-nt matches to seed region (7mer-m8 and 7mer-A1). The third is 8-nt matches to seed region (8mer). The improvement of TargetScans offers users to obtain more confident miRNA target lists.

miRanda

miRanda [145] is the open-source software for miRNA target sites prediction. The target prediction pipeline of miRanda contain three steps. In the first step, the miRNA/mRNA duplex is identified in the 3'UTR region of mRNA. G:U pairs and indels between the duplex are allowed. In the second step, the thermodynamic stability of the miRNA/mRNA duplex is computed. The third step is filtering the predicted target site by the cross-species conservation of target sites. In mammalian, the target sites of human, mouse and rat are used (target conservation>=90%). In fish, zebra fish and fugu are used (target conservation>=70%). miRanda offers users to define the cut-off value for some parameters such as score threshold, energy threshold and the penalty of indels.

PicTar

PicTar [146] is designed to identify common targets of miRNAs. The set of coexpressed miRNAs are used for finding the target locations in the multiple alignments of 3' UTR of mRNAs. PicTar first searches perfect seed which is defined as 7-nt continuous matches starting at position 1 or 2 from the 5' end of miRNAs. Then, predicted target sites are filtered according to the MFE (minimum free energy) of the miRNA target duplex and whether these sites locate into overlapping regions in aligned orthologous sequences. Remained target sites are called as anchors. HMM maximum likelihood score are used for ranking the user-defined minimum number of anchors. The score is calculated by considering all segmentations of the target sequence into target sites and background, thus accounting for the synergistic effect of multiple targeting sites for a single miRNA or varius miRNAs co-regulatinh the same mRNA.

PITA

PITA [147] is the first prediction tool which implements the target site accessibility into identifying the target sites of miRNAs. This concept is observing the difference between the energy of miRNA/mRNA duplex and the energy of unfolded the secondary structure of target region. The energy of miRNA-target duplex and target region are calculated by Vienna RNA packages. The gained energy from miRNA-target duplex is computed by RNAduplex. The folding energy of the target region is computed by RNAfold. The seed match in PITA is defined as beginning from the second position of 5' end of miRNA with 6-8 continuous perfect matches excluding G:U pair. One wobble pairing (G:U pair) is allowed if the length of continuous matches is equal or larger than 7 nt.

RNAhybrid

RNAhybrid [148] is the miRNA target sites prediction tool which does not use the concept of seed match. It identifies all possible MFE in miRNA/mRNA duplexes by using the dynamic programming technique. The length of bulge and internal loops of the miRNA-target duplex are set to be smaller than 15 nt. The most important features of RNAhybrid is that several statistical issues are done such as the length normalization of binding energy, the significance of individual target prediction and modeling multiple binding sites by Poisson approximation. RNAhybrid is available for download and as a Web tool on the Bielefeld Bioinformatics Server (http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/).

3.3 Tool and web servers for analyzing NGS small RNA sequencing data

miRDeep

miRDeep [8] is the first stand-alone package which is designed to identify novel miRNAs from sequencing data generated by next-generation sequencing technology. miRDeep first aligns the sequencing reads with reference genomes. Then, the reads which have multiple genomic loci or mapped to rRNA, tRNAs, scRNA, snRNA and snoRNA are removed. Remained reads are used for finding potential miRNA precursors. Probabilistic scoring systems are applied to each potential precursor. miRDeep can be downloaded at http://www.mdc-berlinde/rajewsky/miRDeep and be executed on the standard Linux machine .

deepBase

deepBase [149] is a database which collected 185 NGS small RNA sequencing data in seven organisms (Homo sapiens, Mus musculus, Gallus gallus, Ciona intestinallis, Drosophila melanogaster, Caenhorhabditis elegans and *Arabidopsis thaliana*). The type of small RNAs which were annotated are nasRNA (ncRNA-associated small RNA), pasRNA (promoter-associated small RNA), easRNA (exon-associated RNA), rasRNA (repeat-associated), miRNA and snoRNA. deepBase also provided the interactive web interface (http://deepbase.sysu.edu.cn) for researchers quickly viewing the annotated small RNA sequencing data .

Geoseq

Geoseq [150] (http://geoseq.mssm.edu) collects deep-sequencing data from various public repositories like GEO (Gene Expression Omnibus) and SRA (Sequence Read Archive) from NCBI and preprocessed these data. The method of Geoseq for dealing with the sequencing data is different with previous studies. It maps the reference sequence against sequencing data instead of mapping them to reference genomes or sequences. Researchers can analyze their own sequencing data against the processing data in Geoseq for identifying differential isoform expression in mRNA-seq datasets and identifying known and novel miRNAs in miRNA datasets.

miRanalyzer

miRanalyzer [11] (http://web.bioinformatics.cicbiogune.es/microRNA/) is the first web server tool for analyzing the next-generation sequencing data in small RNAs. Before uploading the sequencing data, the researchers need to merge the same reads to a unique one and counting their copy numbers (expression level). After running the datasets at miRanalyzer web server, researchers can obtain the analyzed results such as the expression level of all known miRNAs in miRBase, predicted novel miRNA lists and all sequencing reads which can be mapped to transcribed sequences (mRNA, ncRNA and rasiRNA). miRanalyzer also provides the target gene lists for all detected miRNAs by using two miRNA target site prediction tools (miRanda and TargetScan).

SeqBuster

SeqBuster [12] (http://estivill_lav.crg.es/seqbuster) is the web-based toolkit to deal with and analyze high-throughput sequencing small RNA datasets. It also offers the stand-alone version to overcome the storage capacity limitations of the web-based tool. It provides raw data preprocessing, miRNA profiling, the analysis of miRNAs variability (IsomiRs), differentially expressed miRNAs discovery and miRNA target sites prediction. For differentially expressed miRNAs discovery, the equation for normalizing the expression level is n = (freq n/sum [freq all seqs]) × scale-value. SeqBuster is the first tool offering the analysis of IsomiRs. For each detected miRNA, it provides the analysis results in 5' end and 3'end trimming, 5' end and 3'end adding and nt-substitution.

mirTools

mirTools [14] (http://centre.bioinformatics.zj.cn/mirtools/) is the web server which allow researchers to do comprehensive analysis through uploading their high-throughput sequencing small RNA data. In mirTools, researchers can process their raw sequencing data, explore the length distribution of reads, classify reads into different categories such as known miRNAs, snoRNA, rasiRNA and coding sequences, identify novel miRNAs, discovery differentially expressed miRNAs between different samples and predict the target genes of miRNAs by using miRanda and RNAhybrid. mirTools also provide the function analysis of miRNA target genes by investigating them in Gene Onotology terms and pathways .

DSAP

DSAP [13] (http://dsap.cgu.edu.tw) is the web server which is designed for analyzing small RNA datasets produced by next-generation sequencing technology. Researchers need to prepare their datasets as a tab-delimited format which contains the unique reads and their expression level. The system flow of DSAP is different with other web tools for analyzing small RNA sequencing data. Other web tools align reads with reference genomes first. Then, they compare the chromosome location of reads with known miRNA or ncRNA location. DSAP directly align reads with miRNA precursors from miRBase and ncRNAs from Rfam. After mapping reads, the information of cross-species distribution of detected miRNAs is provided. DSAP also provides two or three sample comparison.

miRExpress

miRExpress [9] is the first stand-alone package which is designed for monitoring the miRNA expression and identifying novel miRNAs in small RNA datasets generated by next-generation sequencing technology. miRExpress combine the known miRNAs from miRBase (users do not handle the miRNAs information by themselves). Researchers can use miRExpress to preprocess their raw sequencing data, observe the length distribution of proceeded reads, monitor miRNA expression through user defined parameters, identify novel miRNA by cross-species miRNAs comparing. miRExpress also provides the alignments between detected miRNAs and reads. Users can find nucleotide modification from this report. miRExprees can be downloaded at http://mirexpress.mbc.nctu.edu.tw/ and be executed on x86 Linux 32 or 64 bit system.

miRNAkey

miRNAkey [10] is a package designed to analyze high-throughput sequencing miRNA data. The system flow of miRNAkey contains serval steps. First is trimming 3' adaptor sequence from 3' end of the reads. Second is mapping the reads to known miRNAs. Thrid is counting the expression level of mapped miRNAs and converting the expression into the normalized RPKM expression index (reads per kilobase prer million mapped reads). Fourth is identifying differentially expressed miRNAs by chi-squared analysis. Final is producing the additional information according to the input data, such as multiple mapping levels and post-clipping read lengths. The important improvement is that miRNAkey developed SEQ-EM algorithm for solve the multiple-aligned-reads problems in the detected miRNAs. miRNAkey is freely available for downloading at (http://ibis.tau.ac.il/miRNAkey).

3.4 miRNA target interaction databases and tools

MAGIA

MAGIA [151] (http://gencomp.bio.unipd.it/magia) is the web-based tool which provides miRNA and gene integrated analysis. Users can start their searching by selecting the organism, the Entrez Gene, Refseq and ENSEMBL gene ID and setting the parameters of three miRNA target site prediction tools (miRanda, TargetScan and PITA). Then, users need to upload the expression profile of mRNAs and miRNAs. Different or the same biological samples are allowed. Then, MAGIA run analysis by systematically combining predicted miRNA target genes and the expression profile of miRNA and mRNA. Pearson and Spearman correlations, mutual information, variational Bayesian model and meta-analysis are used for the measure of miRNA-target interactions. The reports of MAGIA provide the bipartite regulatory network which gives the straightforward view of miRNA-target interactions. The useful links in the network are also offered such as miRNA-disease information from miR2Disease, gene expression level in Atlas databases and the text-mining searching link to PubFocus and EbiMed (miRNAs and genes as keywords).

miRGator

miRGator [152] (http://miRGator.kobic.re.kr) is an integrated repository which collects the information of miRNAs, miRNA-associated gene expression, predicted miRNA target sites, miRNA-disease information and genomic annotation in human. The expression profiles of gene in miRGator have two types. One is miRNA knockout. Another is miRNA overexpressed. For miRNA target sites, both known (validated) and predicted miRNA-target interactions are collected. The relationships between miRNAs and disease are obtained from literature. miRGator integrates these collected data and creates various specific relationship such as differentially expressed miRNAs against differentially expressed miRNA target genes and differentially expressed miRNAs against disease-associated miRNAs. miRGator uses the network representation to give the results of different relationships. Moreover, the gene (differentially expressed or miRNA target) annotations are also provided through the information gene ontology (GO) and pathways (KEGG).

TarBase

TarBase [153] (http://www.diana.pcbi.upenn.edu/) is the first database which collects experimental validated miRNA targets. TarBase collects the sets of positive and negative miRNA targets in human, mouse, fruit fly, worm and zebrafish. Positive and negative miRNA targets are judged by detecting the sufficiency of the target site to induce translational repression or cleavage. TarBase also provides the functional links to several databases such as Gene Ontology (GO) and UCSC Genome Browser.

miRecords

miRecords [154] (http://miRecords.umn.edu/miRecords) is an integrated database for miRNA-target interactions. miRecords collects 1135 experimental verified miRNA-target interactions between 301 miRNAs and 902 target genes in seven animal species. miRecords also predicted miRNA targets which are produced by using 11 developed miRNA target site prediction software. At the website of miRecords, user can find the miRNA-target interactions by the miRNA name, the RefSeq accession, the Entrez Gene ID or the gene name of target gene searching.

miR2Disease

miR2Disease [155] (http://www.miR2Disease.org) is the resource for the relationship between miRNA target genes and human disease. miR2Disease collects 1939 relationships between 299 human miRNAs and 94 human diseases. Users can search the relationship of miRNA s and diseases by the miRNA ID, the disease name and the gene name through the web interface of miR2Disease. Moreover, the submission page is provided for researchers submitting the miRNA-disease relationship which is not found in miR2Disease.

miRTarBase

miRTarBase [156] (http://mirtarbase.mbc.nctu.edu.tw/) is the repository which collects 3576 experimental validated MTIs (miRNA-target interactions) between 657 miRNAs and 2297 target genes in 17 species. There are several experimental protocols which can give strong or weak evidence for MTIs. Strong experiment evidence contains Luciferase reporter assay, qPCR (quantitative polymerase chain reaction) and western blot. Weak evidence contains pSILAC and cDNA microarray. pSILAC is designed to high-throughput screening the expression level of proteins when the miRNA is expressed or not. cDNA microarray is used for detect the mRNA expression when the miRNA is present or not. miRTarBase also provides the functional analysis of these verified miRNA target genes by doing Gene Ontology (GO) and pathway (KEGG) enrichment annotations.

starBase

starBase [157] (http://stabase.sysu.edu.cn/) is a database which contains miRNA-mRNA interaction from Argonaute CLIP-Seq and Degradome-seq data. Argonaute CLIP-seq and degradome sequencing (Degradome-seq) which are high-throughput technologies are applied to identify the interaction sites between miRNAs and mRNAs in animals and plants, respectively. 21 CLIP-seq and 10 Degradome-Seq experiments from six organisms are used to identify 1 million Ago-binding clusters in animals and 2 million cleaved target clusters in plants. In animals, TargetScan, PicTar, miRanda, PITA and RNA22 are applied to identify the target sites in detected Ago-binding clusters. Approximate 400000 miRNA-mRNA interactions are identified between 1348 miRNAs and 26296 genes. In plants, approximate 6600 miRNA-target regulatory relationships in 25579 genes and 856 miRNAs are identified by using the CleaveLand program. User can easily obtain various results such as mapped reads, predicted and known miRNA targets, ncRNAs, protein-coidng genes, target-peaks and target-plots at the web site of starBase. Function analysis of miRNA target genes are offered by combining pathways (KEGG and BioCarta) and Gene Ontology (GO). starBase also provides the interface to allow users to predict miRNA target sites in defined sequencing clusters in animals or plants.

3.5 Transcriptional regulation related database and tools

miRStart

miRStart (http://mirstart.mbc.nctu.edu.tw/) is the repository of human miRNA TSSs (transcription start sites). Three published experimental evidences are used for identifying TSSs of miRNAs. The first is Cap-analysis gene expression (CAGE) tags [158]. Cap-analysis gene expression (CAGE) tags are ~20 nts sequences derived from the 5' terminal of cDNA. CAGE tags can be generated using biotinylated cap-trapper with specific linker sequence to ensure the sequence after 5' cap of cDNA was reserved. The CAGE tags contain the first base of 5' terminal sequence, that is, the transcription start site of RNA polymerase II transcripts. The second is ChIP-Seq of histone methylation [159]. ChIP-Seq, a massive parallel signature sequencing technique, offers a high-resolution profiling of histone methylations in the human genome. For example, H3K36me3 and H3K20me1 are associated with actively transcribed regions and support powerful evidence for TSS identification. The third is illumina Solexa tags from DBTSS [160]. Illumina Solexa tags were derived using a new-generation and high-throughput sequencing technology, which DNA templates are immobilized on a special surface fluorescently labeled nucleotides with specific enzyme. The genomic position of Solexa tags in the DBTSS could be directly mapped to the upstream flanking region of intergenic miRNA precursors to support evidence for the detection of miRNA TSSs. The SVM (Support Vector Machine) model is used for incorporating three evidences to identify high-confidence TSSs of miRNAs. Moreover, ESTs (Express short tags) are also provided for reconfirmation.

TRANSFAC

TRANSFAC [161] is a database which collects transcription factors and their binding sites in various organisms. TRANSFAC collects the sequences of TFBS (transcription factor binding site) for each transcription factor (TF). Then, for each TF, the binding profiles called as PWMs (Position Weight Matrix) are generated. For each PWM, the consensus sequence is generated based on IUPAC code. The definition of the consensus sequence is that in the possible type of nucleotides at each position. For example, the consensus is "TACNAYACTTG" from 5'end to 3'end. The nucleotide at position 5 is A (not T, G or C). The nucleotide ate position 6 is C or T (not A or T). The external database annotations of TFs are also provided such as Genbank, EMBL, SWISSPORT and PIR. Users can use the binding profiles to scan whether TFBSs locate in the promoter sequences. TRANFAC is only for economical using now.

JASPAR

JASPAR [162] (http://jaspar.genereg.net) is the open-access database which collects 457 non-redundant matrix profiles. These profiles represent the DNA-binding sequences of transcription factors and other protein-DNA interacting patterns. These profiles are generated by extracting from Chip-seq and Chip-chip experiments. Chip-seq and Chip-chip offers high-throughput and whole genome screening protein-DNA interacting sites. JASPAR also collects 840 profiles which are found by literature searching and other databases such as PBM, PBM_HOMEO and PBM_HLH. Users can download the binding profiles for checking the relationships between TFs and their interested sequences (the promoter of the gene). Users also can use the web interface to easily reach this aim. JASPAR also provides the standardized system for the classification of TFs and grouping into families based on the similarity of binding profiles. User can find easily the relationships in different TFs by this system.

MATCH

MATCH [163] is a tool which is designed for searching putative transcription factor binding sites in DNA sequences. MATCH constructed the scoring system and threshold profiles for each transcription factor binding profiles (PWMs). For the scoring system, the matrix similarity score (MSS) and the core similarity score (CSS) are used for measuring the degree of similarity between the DNA sequences and the matrix. The range of MSS and CSS is 0.0 to 1.0. The definition of the core of the matrix is the first five most conserved continuous positions in the matrix. For the threshold profile, the different cut-off values of two scores are set for each matrix. Three threshold profiles are supported for different purposes. They are for minimizing false positive (over-prediction error), minimizing false negative e (under-prediction error) and minimizing the sum of both errors. MATCH also offers users to transform their own binding sequences into the PWM and construct the default threshold profile. The public version of MATCH can be free downloaded at http://www.gene-regulation.com/pub/programs.

3.6 Other related resources

Gene Ontology

Gene Ontology (http://www.geneontology.org/) is the repository for the annotation of gene products properties. Gene Ontology (GO) contains three major domains. The first is cellular component (cc) which represents the part of a cell or gene's extracellular environment. The second is molecular function (mf) which means the elemental activities of the gene product at the molecular level. The third is biological process (bp) which represents the operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units (cells, tissues, organs and organisms). The tree structure is used to define the scope of each GO term. The node at top of the tree is more global and contains more involved genes. The node at bottom of the tree is more specific and contains fewer involved genes. For example, cell cycle is more global and regulation of meiotic cell cycle is more specific. The description of each GO term and the the tree structure of GO terms can be freely downloaded at the website. The GO annotated gene sets in various species are also available.

KEGG

KEGG [15] (Kyoto Encyclopedia of Genes and Genomes) is the bioinformatics repository including molecular interaction networks (pathways and complexes), the information of genes and proteins, chemical compounds and chemical reactions in various organisms. There are over 6000 drawn pathways in KEGG. The pathway is the protein interaction network which contains three types of interactions such as enzyme-enzyme reaction, direct protein-protein interaction and transcription factors with their target genes. The pathways are categorized into seven classes such as metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human disease and drug development. Users can easily find the roles of their interested genes by observing the function of the pathways which these genes involve. KEGG (http://www.genome.jp/kegg/) also offers users to download the data.

Vienna RNA Package

Vienna RNA package is the free and stand-alone software for RNA secondary structure prediction and comparison. Vienna RNA package can offer users to: 1. predict minimum energy RNA secondary structures; 2. compare RNA secondary structures; 3. predict hybrid structure of two structures and possible hybridization sites; 4. predict the consensus RNA secondary structure from several aligned sequences; 5. draw RNA secondary structures; 6. predict the RNA-RNA interaction sites by using accessibilites. Users can also use these function at Vienna RNA WebServers (http://rna.tbi.univie.ac.at/).

Rfam

Rfam [164] (http://rfam.sanger.ac.uk/) is the repository which collects 1446 RNA families. RNA families contain three function classes such non-coding RNAs, structured cis-regulatory elements and self-splicing RNAs. Rfam offers users to browse the multiple sequence alignments, consensus RNA secondary structures and covariance models (CMs) for each RNA family. Users can also search RNA families by keywords, their interested sequence sets, taxonomy ID and the classes of miRNAs (snRNA, miRNA, cis-regulatory elements…etc).

RNALogo

RNALogo [165] (http://ranlogo.mbc.nctu.edu.tw/) is a web-based tool for generating the graphical representation of RNA consensus secondary structure. Users can upload their interested sequence sets (aligned or not aligned). Then, RNALogo gives the reports which contain conversation of each position by different font size and different colors for the conversation in base-pairing. Users can easily find the key regions of their interested sequence sets from these graphical reports. Moreover, RNALogo supports the preprocessed reports for known regulatory RNAs in Rfam. Users can compare their interested sequence sets with known regulatory RNAs.

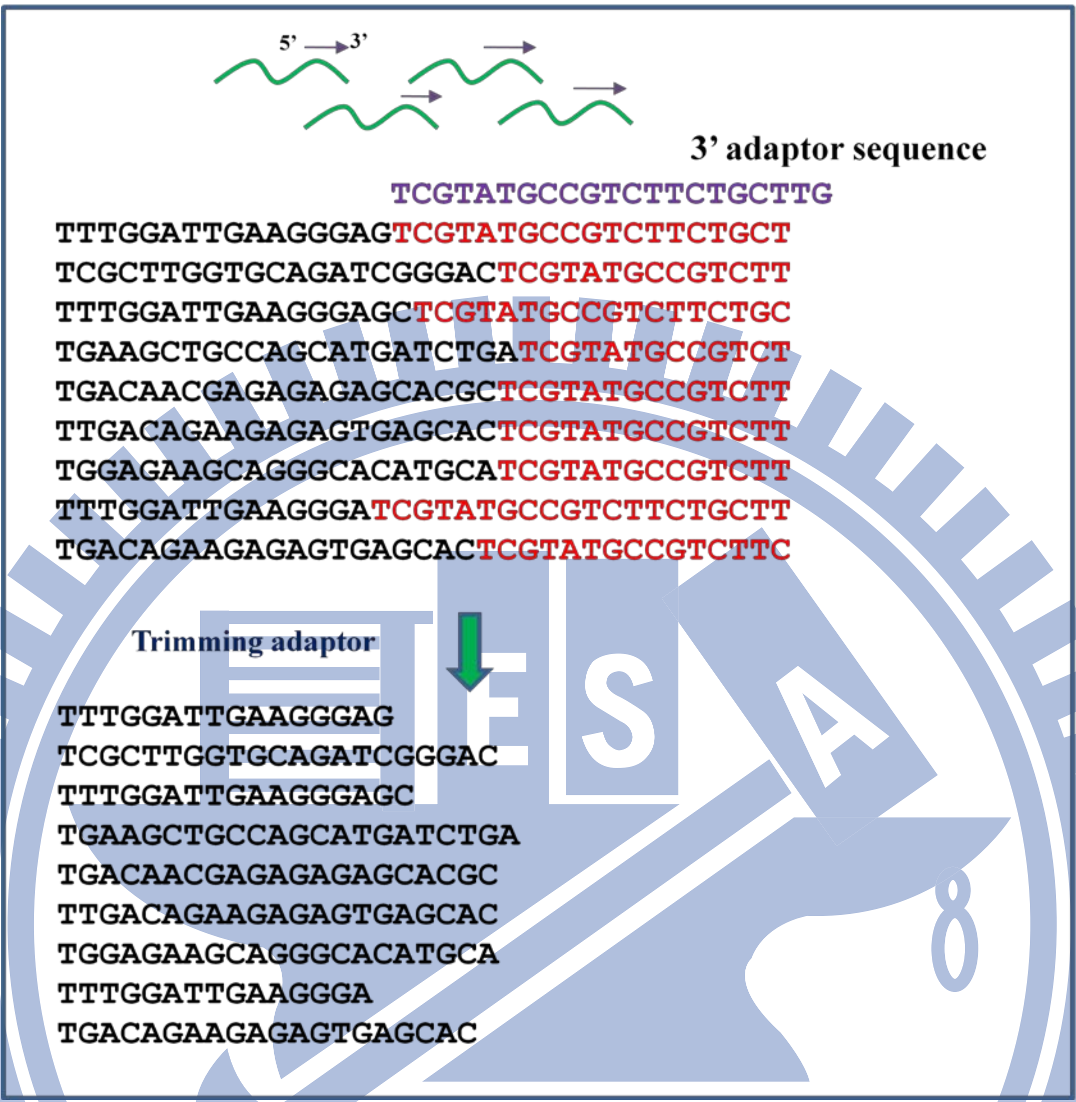
4. Materials and Methods

4.1 The system flow of profiling small RNA expression

For monitoring the expression of small RNA, there are four steps. First step is trimming adaptor sequence from sequencing reads. Second step is aligning trimmed reads with the reference genomes or small RNA database. Third step is calculating the expression level of various kinds small RNA. The fourth is normalizing the expression of small RNA and finding differential expressed small RNAs.

Trimming adaptor sequence from sequencing reads

In the first step, all identical sequencing reads are merged into a unique read and counts each unique read. Then, each unique read is checked to determine whether it contains a full or a partial adaptor sequence at 3' end. The length of reads is 36~40 nt (illumia and ABI SOLiD platforms) and the length of small RNAs is 18~25 nt. Therefore, the reads have full or partial adaptor sequence at 3' end (The length of 3' adaptor is 18~22 nt). In checking the full adaptor sequence, if the adaptor sequence is in the middle or at the beginning of the read, then the read is removed. If the adaptor sequence is at the end of the read, then the adaptor sequence is trimmed from the read. In checking the partial adaptor sequence, the last bases of the 5' adaptor are used as a probe to match the first bases of the reads. The first bases of the 3' adaptor are used as probes to match the last bases of the reads. If the sequence identities of the matched regions are greater than 70%, these regions are eliminated (Figure 4).



3' adaptor sequence

TCGTATGCCGTCTTCTGCTTG

```
TTTGGATTGAAGGGAGTCGTATGCCGTCTTCTGCT
TCGCTTGGTGCAGATCGGGACTCGTATGCCGTCTT
TTTGGATTGAAGGGAGCTCGTATGCCGTCTTCTGC
TGAAGCTGCCAGCATGATCTGATCGTATGCCGTCT
TGACAACGAGAGAGAGCACGCTCGTATGCCGTCTT
TTGACAGAAGAGAGTGAGCACTCGTATGCCGTCTT
TGGAGAAGCAGGGCACATGCATCGTATGCCGTCTT
TTTGGATTGAAGGGATCGTATGCCGTCTTCTGCTT
TGACAGAAGAGAGTGAGCACTCGTATGCCGTCTTC
```

Trimming adaptor

```
TTTGGATTGAAGGGAG
TCGCTTGGTGCAGATCGGGAC
TTTGGATTGAAGGGAGC
TGAAGCTGCCAGCATGATCTGA
TGACAACGAGAGAGAGCACGC
TTGACAGAAGAGAGTGAGCAC
TGGAGAAGCAGGGCACATGCA
TTTGGATTGAAGGGA
TGACAGAAGAGAGTGAGCAC
```

Figure 4. The flow of trimming 3' adaptor sequence from the sequencing reads

Aligning trimmed reads with the reference genomes or small RNA database

In the second step, each read is aligned with the reference genomes or small RNA database (no available reference genomes). If the reference genomes are available, the reads which are perfectly mapped to the genomic sequence are retained. The reads which have multiple chromosome location are removed if the number of location is more than the defined threshold. The threshold is set according to the maximum chromosome locations of mature miRNAs of each species (data are extracted from miRBase 16.0). For example, the maximum chromosome location of Human is 11. The threshold is set as 11. In *Arabidopsis*, it is set as 7. For the sequencing data which do not have available reference genome, the sequencing reads are aligned with small RNA database such as Rfam and miRBase. The perfectly mapped reads are retained for further analysis.

Calculating the expression level of various kinds small RNA

In the third step, the expression level of small RNA is calculated by combing the information of chromosome location of the reads and known small RNAs. Before monitoring the expression level of small RNA, the reads which are mapped to rRNAs, tRNAs, snoRNAs and snRNAs are removed first. The information of these non-coding RNAs are from Rfam and NCBI. In animals, the expression of the miRNA is computed by counting the reads which locate in the region of known miRNA. The region of miRNA is defined as the flanking three-nucleotide region at 5' and 3' end of the miRNA. For example, the mature miRNA of hsa-let-7a-1 is located at chr9: 96938244~96938265 [+]. The region of miRNA is 96938241~96938268. The reads which are located in this region are considered as the expressed sequence of hsa-let-7a-1 (Figure 5). The three-nucleotide is defined by pre-analyzing the global NGS data in different species. Table 3 lists the number of miRNAs and reads and the GEO accession IDs for analyzing the shifting bases of miRNAs in various species. After pre-analyzing NGS data, three-nucleotide shifting for miRNAs can cover more 90% reads in different organisms (Figure 6).

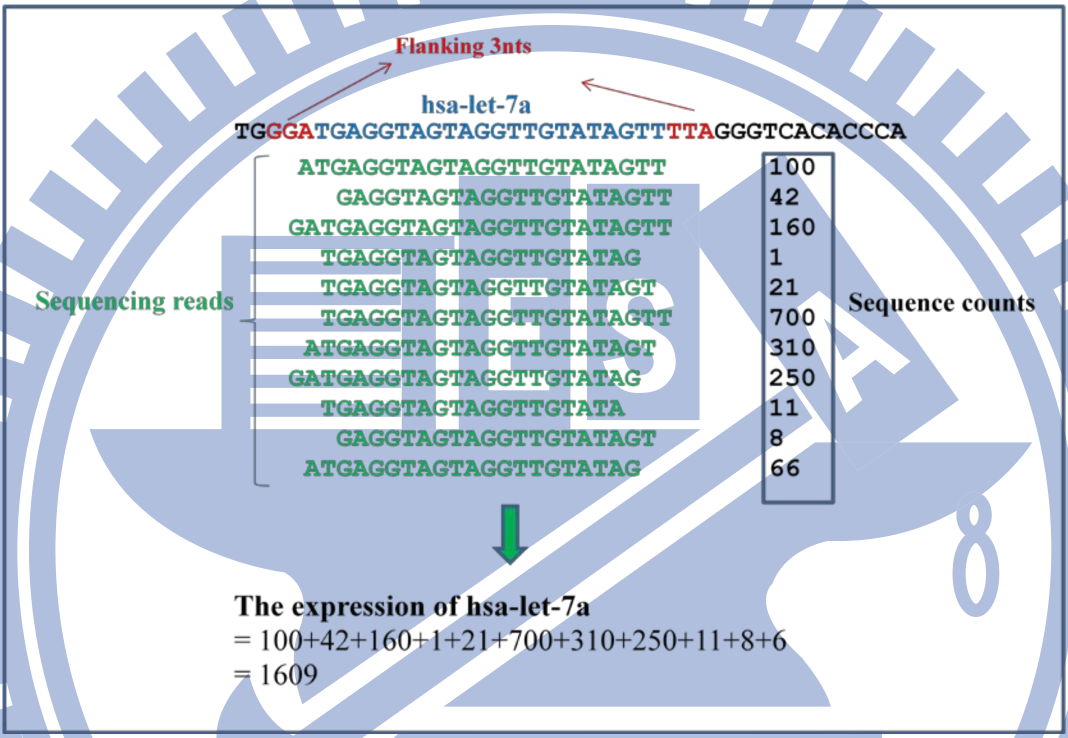


Figure 5. How to monitor the expression level of the miRNA in sequencing data

Table 3. The number of miRNA and reads and GEO accession IDs for analyzing the shifting bases of miRNA in various species

Species	# of miRNA	# of reads	Experiments accession ID (GEO)
Human	774	9433	GSE13370,GSE16579, [17]
Mouse	625	12243	GSE9306,GSE16579,GSE12633,GSE12074,GSE12075,GSE12521,GSE7414
Dog	219	1263	GSE10825
Chicken	489	2666	GSE10686,GSE15513
Drosphilia	161	3043	GSE15378,GSE11086,GSE10794,GSE12840,GSE11019,GSE9389,GSE13679,GSE10515,GSE11624,GSE14849,GSE15898,GSE7448,GSE15899,GSE14488,GSE15137,GSE10790,GSE15897
C.elegan	179	3236	GSE11738,GSE15169,GSE17787,GSE17153,GSE11736,GSE5990
Arabidopsis	215	2388	GSE13605,GSE10036,GSE12037,GSE10180,GSE14696,GSE11094,GSE14695,GSE5343,GSE16971,GSE6478,GSE6682,GSE14694,GSE11007,GSE5228,GSE3008
Rice	416	1825	GSE11974,GSE16350,GSE11014,GSE10523,GSE7107,GSE16248,GSE13152,GSE12317

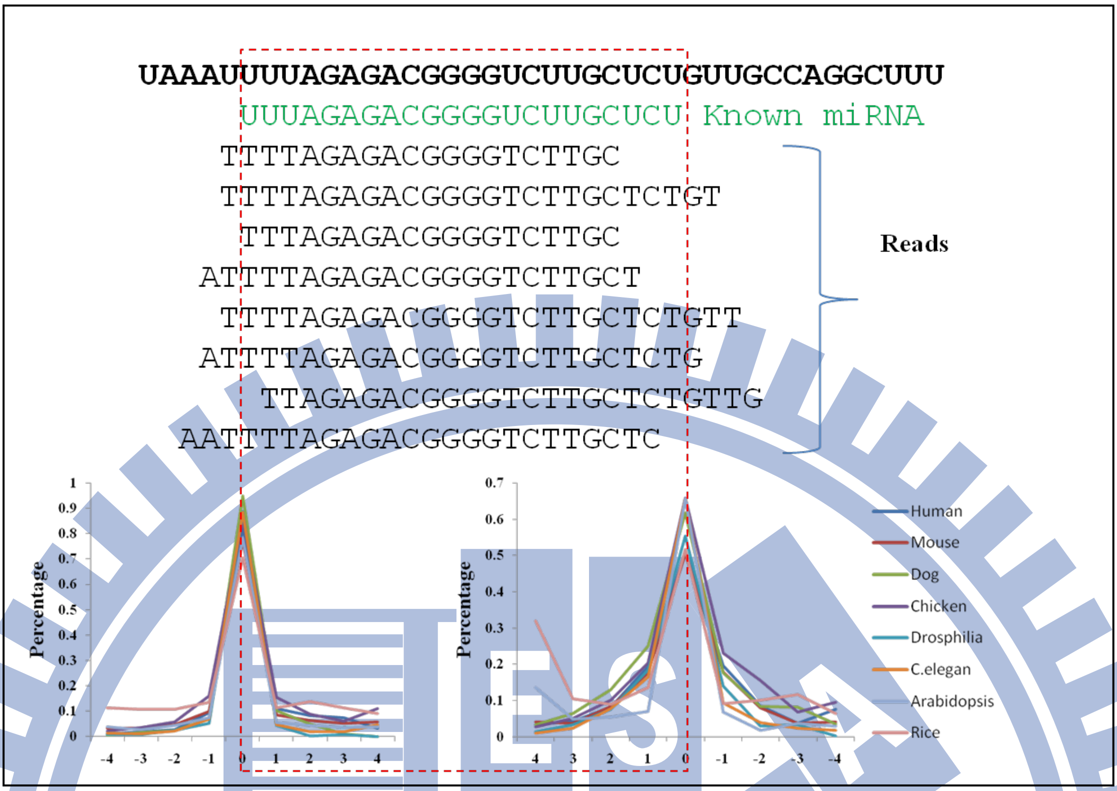


Figure 6. The distribution of shifting bases of miRNAs in varius species

In plants, various kind of small RNAs such as tasiRNA, nat-siRNA and rasiRNA are needed to be profiled excluding the miRNA. For profiling miRNA expression, the processes are the same with animals. For tasiRNAs, the genomic locations of tasiRNAs are obtained from ASRP (*Arabidopsis* Small RNA Project) (http://asrp.cgrb.oregonstate.edu/). The definition of region of tasiRNA is the same with miRNAs (flanking three-nucleotides). The reads which locate in the region are summed to present the expression level of tasiRNA. Nat-siRNAs contain cis- nat-siRNAs and trans- nat-siRNAs. The *cis*-natural antisense transcripts are identified by analyzing the genomic loci of genes from TAIR 9.0 (http://www.arabidopsis.org/), thus the length of overlap regions are greater than 30nts were extracted. The *trans*-natural antisense transcripts are also identified by alignment analysis of all gene transcripts, thus the length of overlap regions, which are with perfect complementarity and greater than 100 nts are extracted. The expression level of nat-siRNAs is calculated by counting the reads which locate in cis- or trans- natural antisense region. In profiling the expression of rasiRNA, the transposons and repetitive elements are obtained from Repbase (http://www.girinst.org/). Inverted and tandem repeats are extracted by using "Inverted Repeats Finder" [166] and "Tandem Repeats Finder" [167] to scan the whole genome, respectively. The reads locate in the dsRNA regions which are formed by transposons, repetitive elements, inverted and tandem repeats are counted as the expression level of rasiRNAs.

Normalizing the expression of small RNA and finding differential expressed small RNAs

In the fourth step, in order to find differentially expressed small RNAs in different experiments, normalization is needed to be done first for each experiment. The quantities of reads which are multiplied one million and divided by adjusted read counts are called as rpm (reads per million).

$$\text{RPM} = \frac{\text{read counts}}{\text{Adjusted read counts}} \times 1{,}000{,}000$$

"Adjusted read counts" indicates the total amount of mapped reads subtracted the amount of reads, which were annotated as rRNAs, tRNAs, snoRNAs and snRNAs. After normalization, differentially expressed small RNAs are found if the fold change is more than 1.5 (up-regulated) or less than 0.67 (down-regulated).

4.2 Novel miRNAs discovery

After profiling known miRNA expressions, there are still some reads which can't be mapped to known miRNAs but can be mapped to genomes. To check these reads whether are the novel miRNAs or not. First, these reads are clustered according to their chromosome location. The maximum distance between two reads is 10 nt. Then, if the distance between two clusters is smaller than 100 nt, these two clusters are merged to one cluster (These two clusters could be miRNA and miRNA*). The flanking regions of these clusters are extracted. In animals, the flanking length is 100 nt at 5' and 3' end. In plants, the length is set as 200 nt. Then, these flanking sequences are folded RNA secondary structures by using RNAfold which identifies stable RNA structures. These RNA secondary structures are filtered based on the characteristics of miRNA precursors. The structures are removed if they are not stem-loop structures (the reads must locate in the stem). The structures are also removed if the ratio of base pairing in the stem region or the number of unpairing bases is not fitted with thresholds. These thresholds are different in each organism and constructed by pre-analyzing the RNA structures of the known miRNAs from miRBase. Table 4 lists the thresholds for the ratio of base pairing and the number of unpairing bases in animals and plants. For example, the thresholds for the ratio of base pairing and for the number of unpairing bases are 0.69 and 6 in human, respectively. There are 1355 human miRNAs and 90% miRNAs are fitted with these thresholds. Therefore, the structures are retained if the ratio of base pairing is greater than 0.69 and the number of unpairing bases is smaller than 6. Finally, the remaining structures are drawn as the figures and checked manually whether the secondary structures are unbranched fold-back (hairpin liked) (Figure 7).

To find more reliable novel miRNAs, the miRNA candidates are checked with cross-species conserved region. The concept of cross-species conserved region is widely used in identifying novel miRNAs. For example, if a known human miRNA locates in human, mouse and rat conserved region, this region may contain a miRNA in mouse and rat. The first appear is identifying novel C. *elegans* miRNAs [168]. Afterward, this concept is applied for discovery miRNAs in human [169] and fly [170]. Recently, 529 novel chimp miRNAs were identified based on this concept [171]. The cross-species conserved sequences of miRNA candidates are extracted from UCSC Genome Browser [172]. Then, RNALogo which is the tool supporting RNA sequence and structure conservation information is applied to observe the conservation of the novel miRNAs. If they are highly conserved, they are more convinced as the novel miRNAs.

Table 4. The thresholds of miRNAs and corresponding coverage rates for each organism

Species	# of miRNAs	the ratio of base pairing	# of unpairing bases	coverage
Animals				
hsa	1355	0.69	6	0.90
mmu	1178	0.69	5	0.91
bta	723	0.63	7	0.90
rno	714	0.71	5	0.91
ppy	638	0.67	4	0.90
bmo	610	0.62	7	0.90
ptr	603	0.67	6	0.90
gga	584	0.67	7	0.90
oan	557	0.71	6	0.90
cin	534	0.64	6	0.91
mml	517	0.72	8	0.90
dre	395	0.73	4	0.90
eca	387	0.71	5	0.92
cfa	325	0.70	5	0.90
dps	302	0.70	5	0.90
sme	271	0.70	5	0.92
tgu	269	0.72	6	0.91
ssc	255	0.73	5	0.90
cel	235	0.70	4	0.90
xtr	207	0.74	4	0.92
dme	201	0.69	5	0.92
dsi	182	0.71	5	0.91
mdo	161	0.72	5	0.91
cbr	150	0.70	3	0.91
crm	149	0.74	3	0.91
aae	140	0.72	4	0.91
cte	133	0.73	4	0.90
tni	132	0.73	17	0.90
fru	131	0.75	4	0.90
ppc	124	0.73	3	0.91
api	123	0.66	5	0.92
cqu	101	0.71	4	0.90
nve	96	0.78	3	0.91
ppa	92	0.75	4	0.91
dgr	88	0.72	4	0.91
ggo	88	0.73	4	0.91
der	84	0.72	4	0.90
dya	83	0.72	4	0.90
dse	82	0.72	5	0.91
dwi	81	0.72	4	0.91
bfl	79	0.75	4	0.91
dan	79	0.74	4	0.91
dpe	78	0.70	4	0.92
sja	78	0.66	4	0.91
dvi	77	0.71	4	0.92
mne	77	0.74	4	0.91
dmo	76	0.69	5	0.91
aga	68	0.72	5	0.91
odi	66	0.69	6	0.91
lgi	64	0.77	3	0.91
ame	63	0.72	3	0.92
age	61	0.73	3	0.93
tca	58	0.76	4	0.91
nvi	51	0.76	3	0.90
lla	50	0.72	3	0.90
sko	45	0.74	5	0.91
spu	45	0.73	5	0.91
dpu	44	0.63	19	0.91
sla	43	0.73	4	0.91

isc	34	0.73	4	0.97
bma	32	0.75	3	0.91
ngi	32	0.75	6	0.91
nlo	28	0.76	6	0.93
csa	27	0.66	6	0.93
xla	22	0.72	5	0.91
hma	20	0.76	3	0.90
lca	17	0.69	3	0.94
aqu	16	0.82	1	1.00
ola	15	0.54	17	0.93
lmi	14	0.74	3	0.93
pbi	11	0.80	3	0.91
ssy	11	0.72	1	0.91
hru	5	0.74	3	1.00
sma	5	0.73	2	1.00
oar	3	0.78	2	1.00
cla	2	0.71	4	1.00
cgr	1	0.82	3	1.00
		plants		
osa	508	0.76	4	0.90
mtr	376	0.72	5	0.90
aly	374	0.80	3	0.90
zma	316	0.76	4	0.90
ppt	277	0.76	2	0.92
ath	241	0.75	4	0.90
ptc	234	0.71	4	0.91
gma	205	0.80	3	0.90
vvi	184	0.76	2	0.92
sbi	147	0.80	3	0.91
cre	85	0.81	2	0.91
smo	64	0.80	2	0.92
csi	63	0.72	3	0.90
rco	63	0.80	2	0.94
bna	48	0.80	3	0.96
aqc	45	0.80	2	0.91
pab	41	0.69	5	0.90
pta	38	0.72	8	0.92
tae	37	0.75	4	0.92
ghr	36	0.81	2	0.94
ahy	32	0.74	3	0.91
sly	30	0.80	4	0.90
bra	23	0.81	2	0.96
bdi	18	0.39	18	0.94
sof	16	0.80	3	0.94
hvu	14	0.66	6	0.93
gso	13	0.81	2	1.00
far	11	0.18	21	0.91
pvu	10	0.76	3	0.90
bol	7	0.81	3	1.00
ctr	6	0.77	2	1.00
ccl	5	0.86	2	1.00
peu	5	0.78	1	1.00
crt	4	0.77	1	1.00
gra	4	0.87	2	1.00
lja	4	0.90	1	1.00
ata	2	0.74	2	1.00
mdm	2	0.80	3	1.00
vun	2	0.95	1	1.00
cpa	1	0.86	0	1.00
gar	1	0.95	1	1.00
ghb	1	0.90	1	1.00
ttu	1	0.86	0	1.00

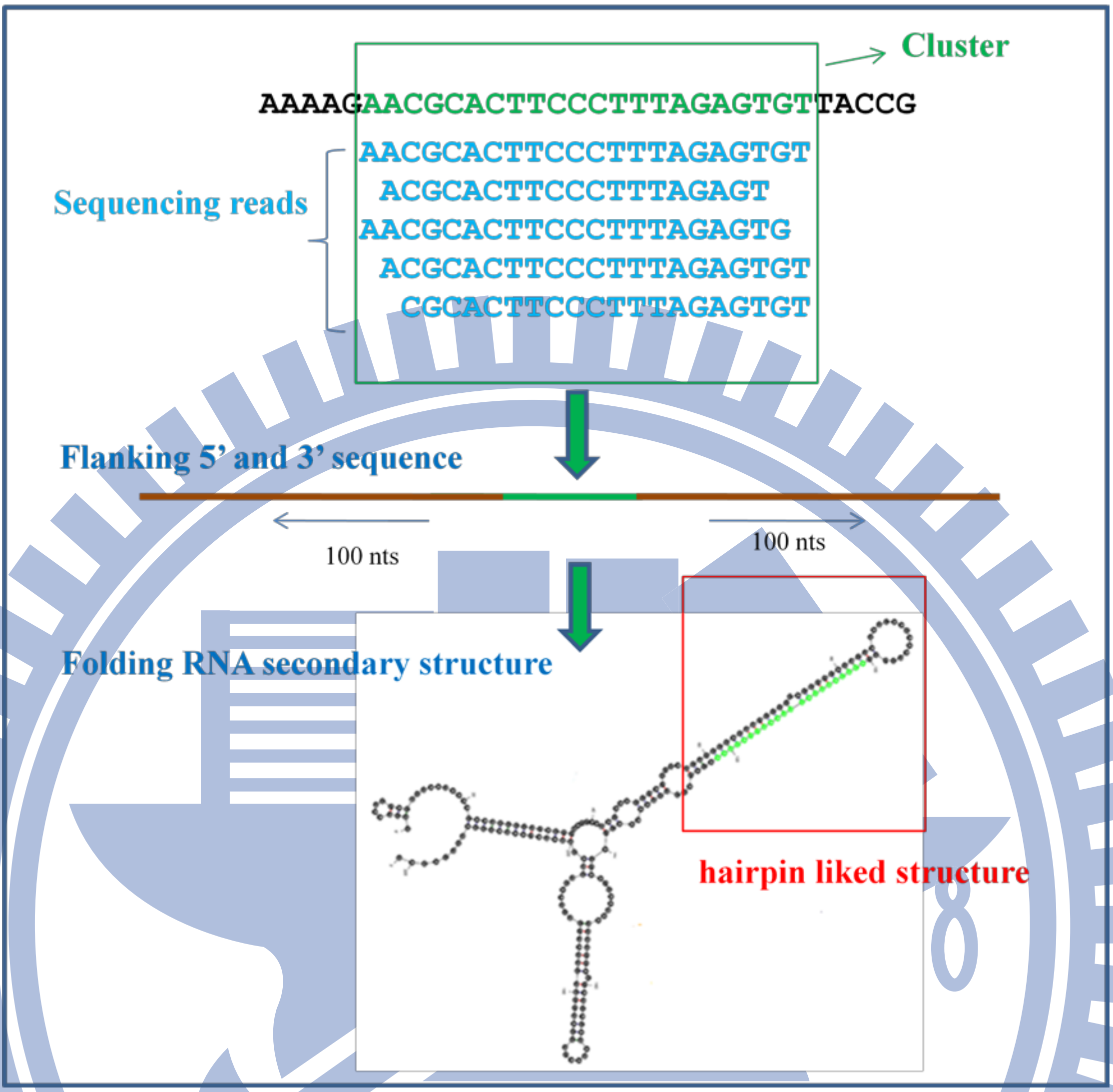


Figure 7. The flow of finding the novel miRNA. It contains clustering reads, flanking the cluster and folding RNA secondary structure

4.3 miRNA and siRNA target genes prediction

For miRNA target prediction in animals, the degree of complementarity between seed region, 2~8 of the miRNA, and 3'UTR of mRNA is the main standard for judging whether is the miRNA target or not. Several miRNA target prediction tools such as miRanda, TargetScans, PicTar and RNAhybrid were developed and widely used. The 3' UTR sequences of mRNA are extracted from GenBank. First, the miRNA and mRNA duplexes are predicted by using miRanda. For miRanda, the targets are identified which the MFE (minimum free energy) between the miRNA and target duplex is smaller than−12 kcal/mol and the miRanda score exceed 140. As the previous studies, many validated miRNA target sites located in cross-species conserved regions. Moreover, conserved miRNAs targeted the same genes. For example, hsa-miR-33a targeted ABCA1 in human and mmu-miR-33 targeted Abca1 in mouse [139-141]. Therefore, TargetScans which can check the miRNA target sites locate in cross-species conserved regions or not is used to filter the candidates from miRanda.

For identifying siRNA target genes in plants, the standard for the siRNA/target duplex is that it needs to be nearly perfect complementary. And the siRNA target sites usually locate at the coding region. The whole transcripts are scanned the complementarity with the siRNAs. The target candidates are filtered according to the sequence complementarity between a siRNA and its target gene, thus the siRNA/target duplex with the sequence complementarity greater than 80% are retained. A series of criteria are designed to select more probable siRNA target genes by modifying the guideline for prediction siRNA targets in plants derived by a previous work [114]. The scoring system is that mismatched pairs or bulges are scored as 1 and G:U pairs are scored as 0.5. The modified criteria are listed as follows. The core region of siRNA ranges from position 3 to 12. Only one bugle (gap) locating in siRNA is allowed and the score of gap is changed from 1 to 2. Nucleotide pairs at position 10 and 11 of siRNA/target duplex must be G:C and A:U pairs. Two continuous mismatch base pairs are allowed. The score of the two continuous mismatch region, which does not contain G:U pairs, is multiplied by 1.5. Three continuous mismatch base pairs are allowed if the mismatch region contains at least two G:U pairs and the score of the region is multiplied by 1.5. More than three continuous mismatch base pairs are not allowed. The score of siRNA/target duplex <= 5.5 are selected as potential siRNA targets.

To reduce the false positive rate of target prediction, the gene expression data of cDNA microarray is used. The function of miRNA is regulating gene expression through degradation of mRNAs or repression of translation. Therefore, if the genes are miRNA direct targets, the mRNA expression level of gene will be down-regulated (degradation of mRNAs) in the cDNA microarray. The high-confidence miRNA targets are constructed based on the criteria of selecting miRNA targets listed as follows. The predicted target genes are retained if the miRNAs are up-regulated and their target genes are down-regulated or the miRNAs are down-regulated and their target genes are up-regulated.

4.5 cis-regulatory elements transcriptional regulation analysis of miRNA

In order to identify cis-regulatory elements of miRNAs, the transcriptional start sites (TSSs) of miRNA genes need to be identified first. There are two kinds of miRNAs. One is intragenic miRNA. Another is intergenic miRNA. For most intragenic miRNAs which are harbored in annotated genes, the regulatory mechanism is coincided with their host genes. It is implied that intragenic miRNAs and their host genes share the common TSSs and express simultaneously. For intergenic miRNAs, the TSSs of miRNA are obtained from miRStart (http://mirstart.mbc.nctu.edu.tw/). miRStart is the database which offers identified TSSs of miRNA in human by combining experimental evidence such as ChIP-Seq of histone methylation, Cap-analysis gene expression (CAGE) tags and illumina Solexa tags from DBTSS.

After identifying the TSSs of miRNAs, the promoter sequences of miRNA are extracted (flanking -3000~+500 according to TSS). For finding transcription factor binding sites (TFBSs) in the promoter sequences, all available PWMs (position weighted matrix) are obtained from TRANSFC and JASPAR which collected experimental validated TFBSs in various species. The TFBSs are scanned in the promoter sequences by applying all PWMs to MATCH which is a searching TFBS tool in DNA sequences (Figure 8).

To reduce the false positive rate of TFBS prediction, the gene expression data of cDNA microarray is used. If the transcription factors (TFs) can directly regulate miRNAs, their expression level will change when the expression level of miRNA change. It is means that the miRNAs are expressed (up or down-regulated) and TFs which regulated them are also expressed (up or down-regulated). According to this rule, the TFs are removed if their expression levels do not be detected or change. The more reliable TFs can be constructed by this method.

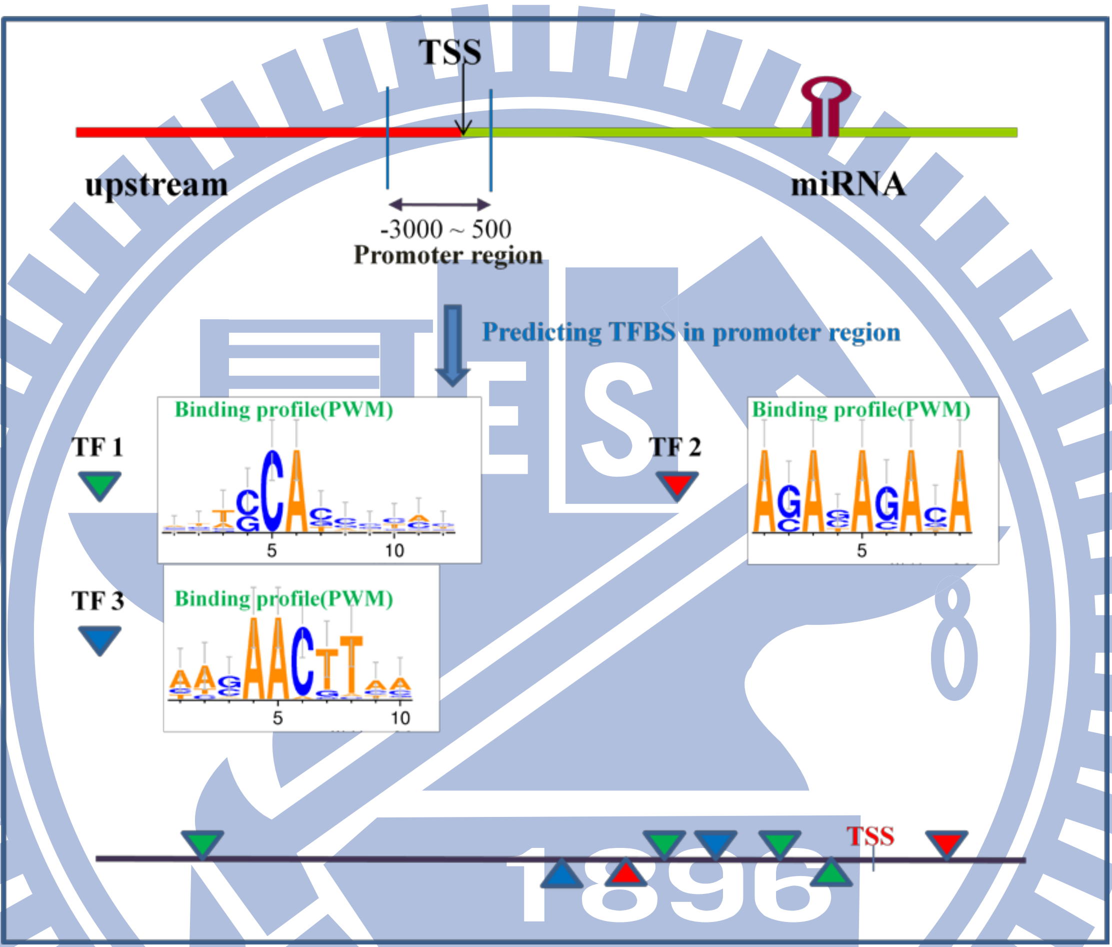


Figure 8. The flow of identifying TFBS in the promoter region of miRNAs. The sequence logos of binding profile of TFs are constructed by weblogo

4.6 Regulation role of miRNA analysis

To find the regulation role of miRNA target genes, the hypergeometric test is applied to the analysis of GO (Gene ontology) terms and pathways. Hypergeometric distribution is one kind of discrete probability distribution which describes the total times of successes after continuous ***n*** draws from the fixed sets without replacement. The formula is following:

$$P[X=k]=\frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

The example which is usually used is drawing black and white balls. ***N*** is the number of white and black balls. ***n*** is the number of draw white and black balls. ***m*** is the number of white balls. ***k*** is the number of draw white balls. The value of this formula means the probability of draw ***k*** white balls. The probability of draw at least ***k*** white balls (cumulative hypergeometric probability) is following:

$$P[X\geq k]=\sum_{x=k}^{m}\frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

The formula of cumulative hypergeometric probability is applied for Gene Onotology (GO) and pathways enrichment analysis (described in the following two sections).

Gene ontology enrichment analysis

In functional analysis of miRNA target genes, the information of GO terms is used. GO is the collection of biological function descriptions. It contain three parts such as Molecular function (MF), Biological process (BP) and Cellular compartment (CC). Each GO term contains the information of biological function and gene sets which involve in this function. For each GO term, the formula of hypergenometric test is applied to checking this biological function is significant or not for miRNA target genes. The meanings of the parameters in the hypergenometric test are following. ***N*** is the number of whole genes having GO terms. ***n*** is the number of miRNA target genes. ***m*** is the number of genes which have selected GO annotation. ***k*** is the number of miRNA target genes which have selected GO annotation (Figure 9). The significant biological function of these miRNA target genes are identified by using this test (The cut-off value is set as <0.01).

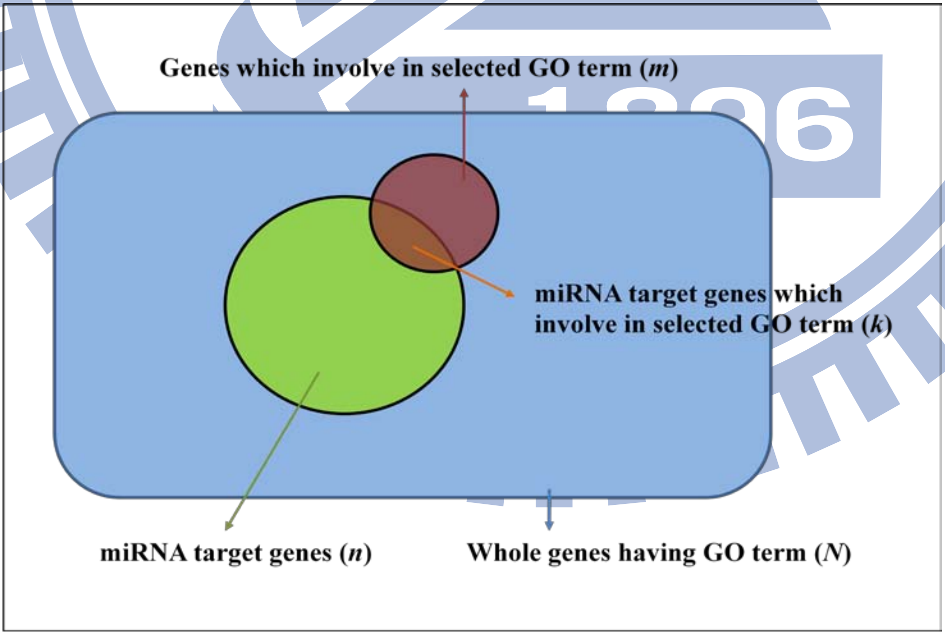

Genes which involve in selected GO term (*m*)
miRNA target genes which involve in selected GO term (*k*)
miRNA target genes (*n*)
Whole genes having GO term (*N*)

Figure 9. The intersection of hypergeomteric test in gene ontology enrichment analysis

Pathway enrichment analysis

To investigate the influence of miRNA target genes in the pathway, the pathway information are collected from KEGG which is a database containing the information of biological systems. The hypergeometric test is used to identify significant pathways which are contained miRNA target genes. The meanings of the parameters in the hypergenometric test are following. ***N*** is the number of genes in all pathways. ***n*** is the number of miRNA target genes. ***m*** is the number of genes in the selected pathway. ***k*** is the number of miRNA target genes in the selected pathway (Figure 10). The significant pathways are identified by the cut-off value (<0.05).

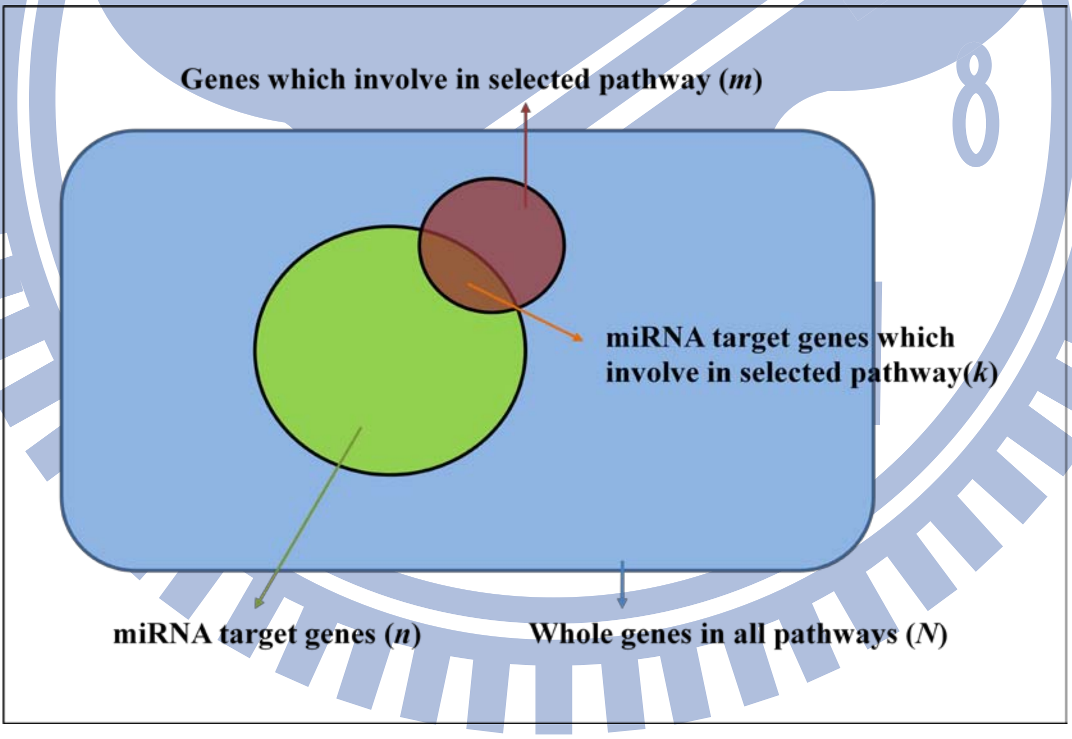

Genes which involve in selected pathway (*m*)
miRNA target genes which involve in selected pathway(*k*)
miRNA target genes (*n*)
Whole genes in all pathways (*N*)

Figure 10. The intersection of hypergeomteric test in pathway enrichment analysis

If the gene expression data of cDNA microarray is available, the gene lists for hypergenometric test are differential expressed genes (containing miRNA target genes). So, the meanings of the parameters are changed. ***N*** is the number of genes in all pathways. ***n*** is the number of differential expressed genes (containing miRNA target genes). ***m*** is the number of genes in the selected pathway. ***k*** is the number of differential expressed genes in the selected pathway. The regulation role of miRNAs and how they affect the process of pathways can be more clearly understood by selecting significant pathways through this way.

5. Results

Case Study 1

To evaluate the workability of the pipeline of analyzing small non-coding RNA from NGS, the public domain data is used. This data (GEO accession ID: GSE19833) is downloaded from GEO (Gene Expression Omnibus) which is a repository collecting array- and sequence-based high-throughput data. This data contains small RNA data from NGS and gene expression data from microarray in human normal and cancer cells [34]. The normal cell is peripheral blood mononuclear (PBMC) cell. The cancer cell lines are K562 and HL60.

For monitoring the expression of miRNAs from NGS, 760 expressed miRNAs are detected in normal and cancer cells. Then, 21 differentially expressed miRNAs are identified by selecting both up- or down-regulated in two cancer cell lines (fold change>=1.5 or <=0.67). Table 5 demonstrates these 21 differentially expressed miRNAs. Among them, only hsa-miR-25 is up-regulated and other 20 miRNAs are down-regulated. In addition to profiling the expression level of known miRNAs, 16 novel miRNA candidates are identified. Figure 11~18 show their RNA secondary structure, chromosome locations and cross-species conservation. Most of them are highly conserved in the stem (mature miRNAs). In 16 novel miRNAs, five of them are opposite strand of known miRNAs in the stem. They are hsa-miR-382 (the second novel miRNA), hsa-miR-1185 (the fifth), hsa-miR-365 (the sixth), hsa-miR-642 (the tenth) and hsa-miR-539 (the fifteenth). Some of these five novel miRNAs are annotated in other species such as mouse, chimp, rat, cow and dog.

Table 5. The lists of differentially expressed miRNAs in two cancer cell lines

	Normalized miRNA expression			Fold change	
miRNA name	PBMC (Normal)	K562	HL60	K562/PBMC	HL60/PBMC
hsa-miR-25	6525.68	10451.74	10090.54	1.60	1.55
hsa-miR-146a	101.70	43.92	13.52	0.43	0.13
hsa-miR-192	3509.45	2314.58	214.53	0.66	0.06
hsa-miR-24	730.75	198.35	472.27	0.27	0.65
hsa-miR-26b	2245.34	1171.69	1373.37	0.52	0.61
hsa-miR-148a	566.51	350.42	206.55	0.62	0.36
hsa-miR-22	897.61	249.83	16.84	0.28	0.02
hsa-miR-34c-5p	67.14	29.75	2.22	0.44	0.03
hsa-miR-27a	427.64	103.43	152.70	0.24	0.36
hsa-miR-320d	55.39	20.78	5.98	0.38	0.11
hsa-miR-106b	1339.49	648.42	531.66	0.48	0.40
hsa-miR-1	685.61	127.51	41.89	0.19	0.06
hsa-miR-181d	49.61	25.03	20.39	0.50	0.41
hsa-miR-151-3p	506.51	309.33	1.55	0.61	0.00
hsa-miR-21	34362.24	15766.16	4010.42	0.46	0.12
hsa-miR-152	278.81	89.26	27.48	0.32	0.10
hsa-miR-361-3p	57.07	2.36	6.65	0.04	0.12
hsa-miR-26a	2246.33	949.26	568.23	0.42	0.25
hsa-miR-101	15079.35	5703.57	6660.54	0.38	0.44
hsa-miR-16	1763.13	757.04	856.12	0.43	0.49
hsa-miR-27b	403.27	166.71	48.98	0.41	0.12

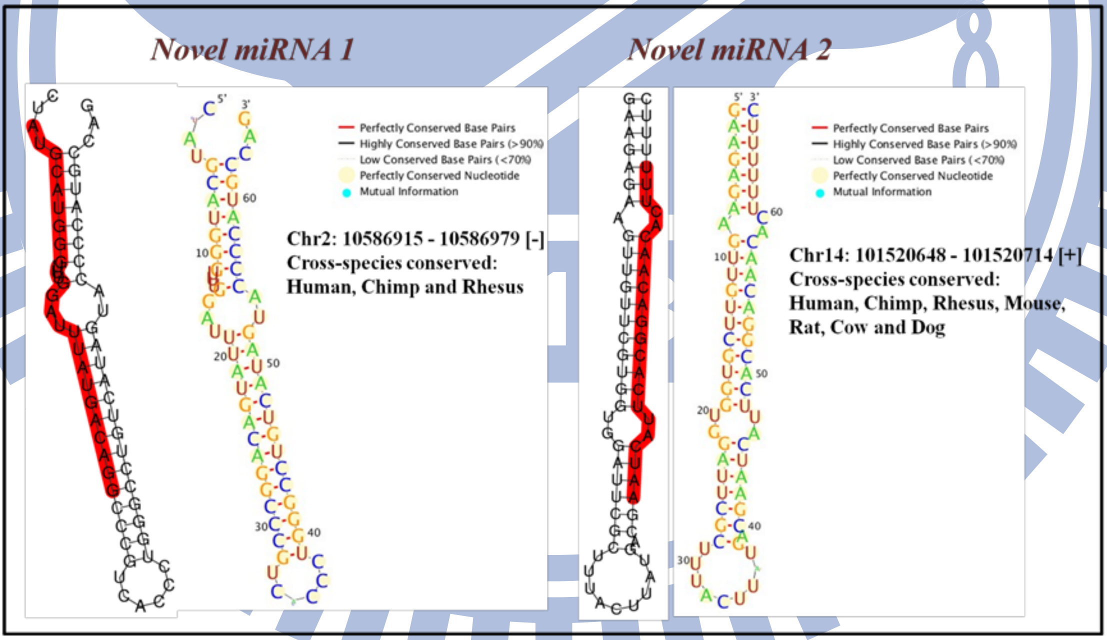

Novel miRNA 1
Novel miRNA 2
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr2: 10586915 - 10586979 [-]
Cross-species conserved:
Human, Chimp and Rhesus
Chr14: 101520648 - 101520714 [+]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Rat, Cow and Dog
5'
3'
10
20
30
40
50
60

Figure 11. The secondary structure and cross-species conservation of novel miRNA 1 and 2 (case study 1)

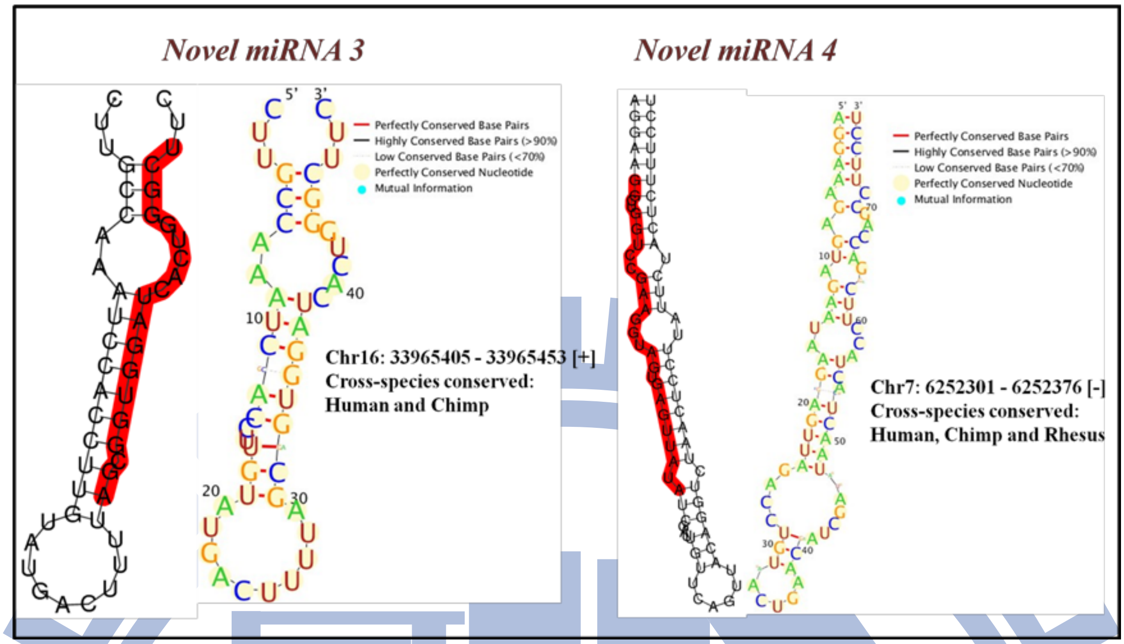

Novel miRNA 3
Novel miRNA 4
5'
3'
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr16: 33965405 - 33965453 [+]
Cross-species conserved:
Human and Chimp
Chr7: 6252301 - 6252376 [-]
Cross-species conserved:
Human, Chimp and Rhesus

Figure 12. The secondary structure and cross-species conservation of novel miRNA 3 and 4 (case study 1)

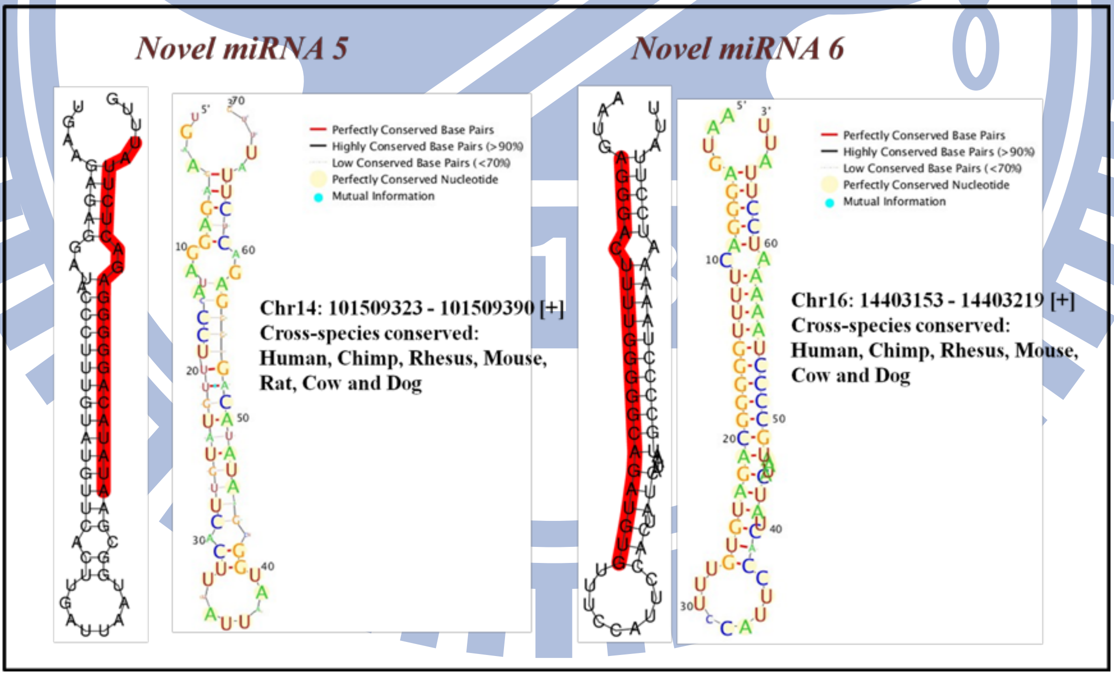

Novel miRNA 5
Novel miRNA 6
5'
3'
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr14: 101509323 - 101509390 [+]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Rat, Cow and Dog
Chr16: 14403153 - 14403219 [+]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Cow and Dog

Figure 13. The secondary structure and cross-species conservation of novel miRNA 5 and 6 (case study 1)

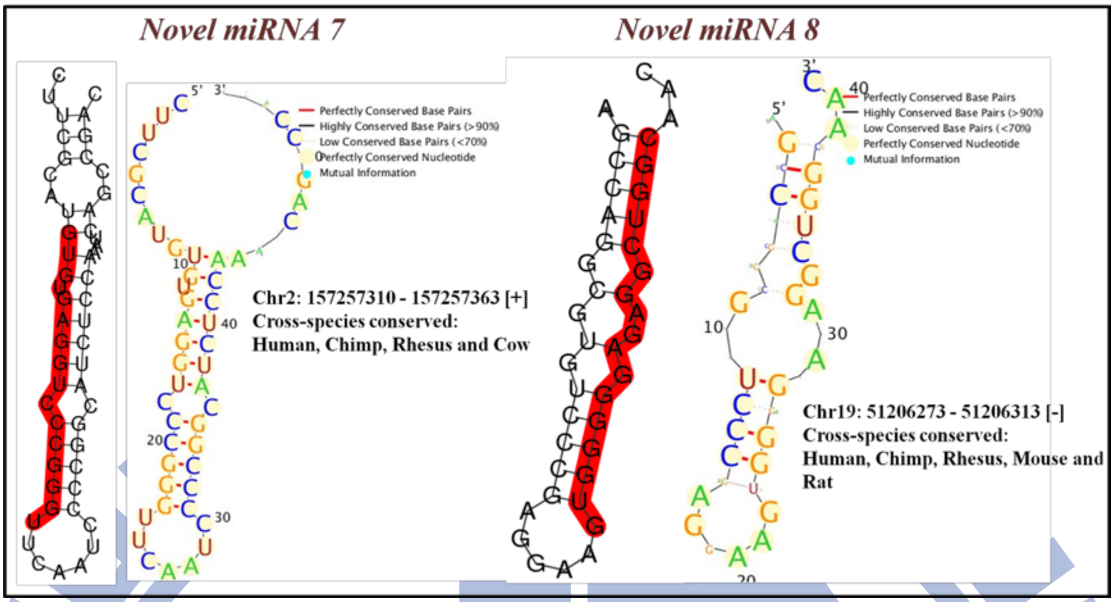

Novel miRNA 7
Novel miRNA 8
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr2: 157257310 - 157257363 [+]
Cross-species conserved:
Human, Chimp, Rhesus and Cow
Chr19: 51206273 - 51206313 [-]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse and
Rat

Figure 14. The secondary structure and cross-species conservation of novel miRNA 7 and 8 (case study 1)

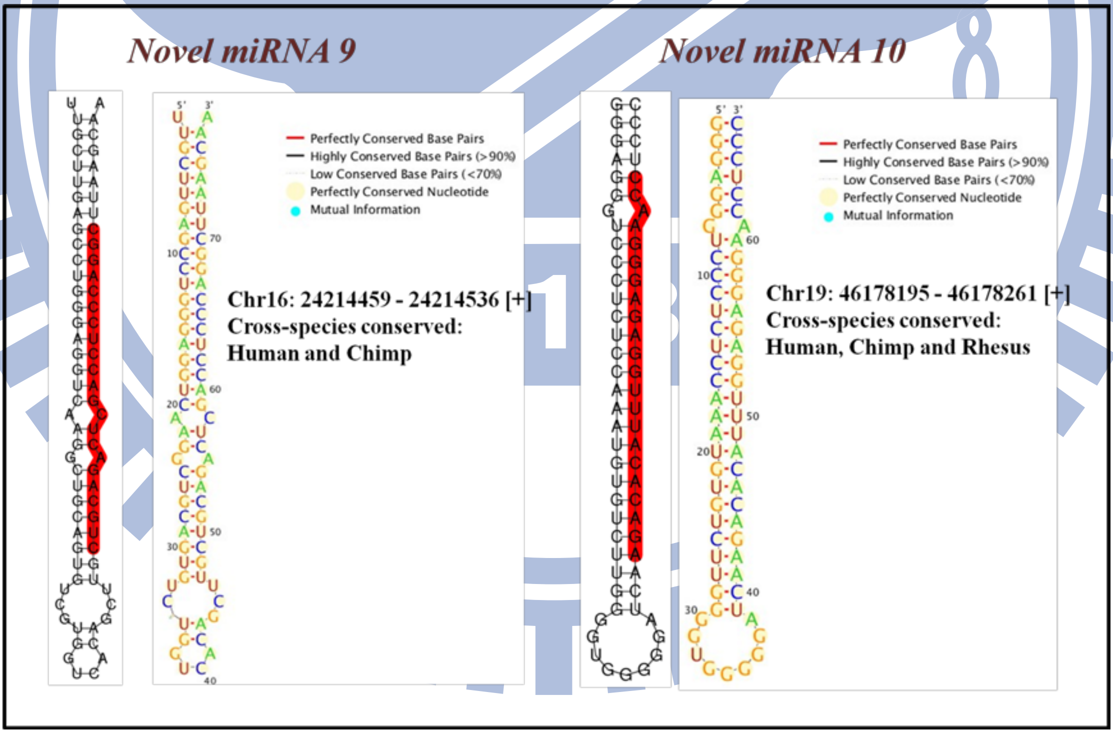

Novel miRNA 9
Novel miRNA 10
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr16: 24214459 - 24214536 [+]
Cross-species conserved:
Human and Chimp
Chr19: 46178195 - 46178261 [+]
Cross-species conserved:
Human, Chimp and Rhesus

Figure 15. The secondary structure and cross-species conservation of novel miRNA 9 and 10 (case study 1)

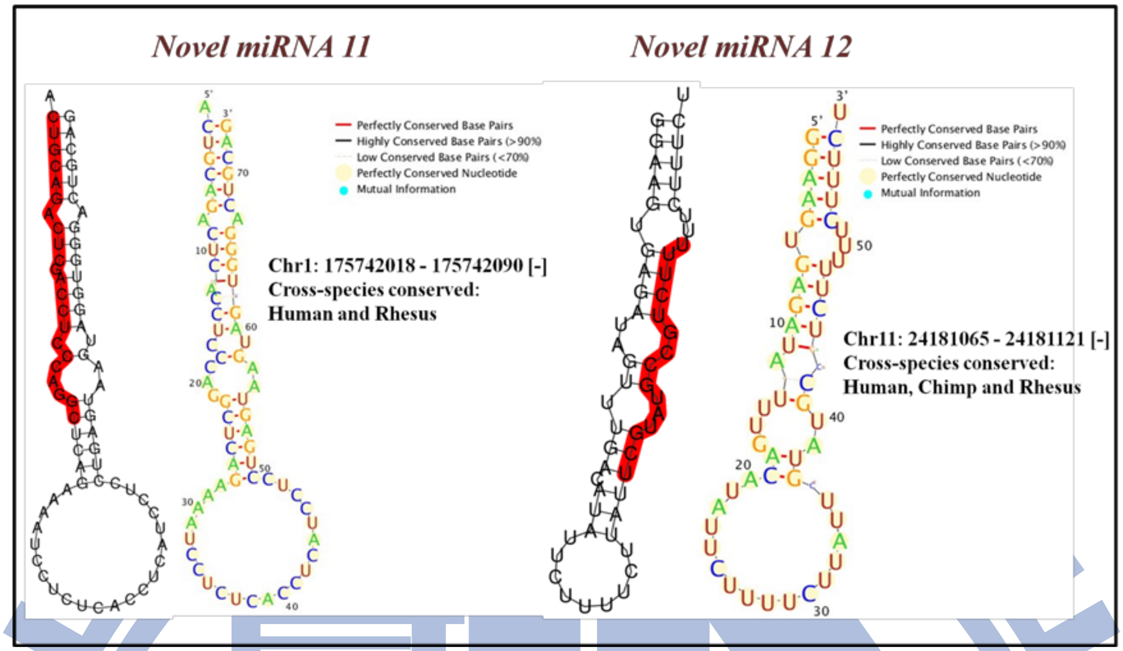

Novel miRNA 11
Novel miRNA 12
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr1: 175742018 - 175742090 [-]
Cross-species conserved:
Human and Rhesus
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr11: 24181065 - 24181121 [-]
Cross-species conserved:
Human, Chimp and Rhesus

Figure 16. The secondary structure and cross-species conservation of novel miRNA 11 and 12 (case study 1)

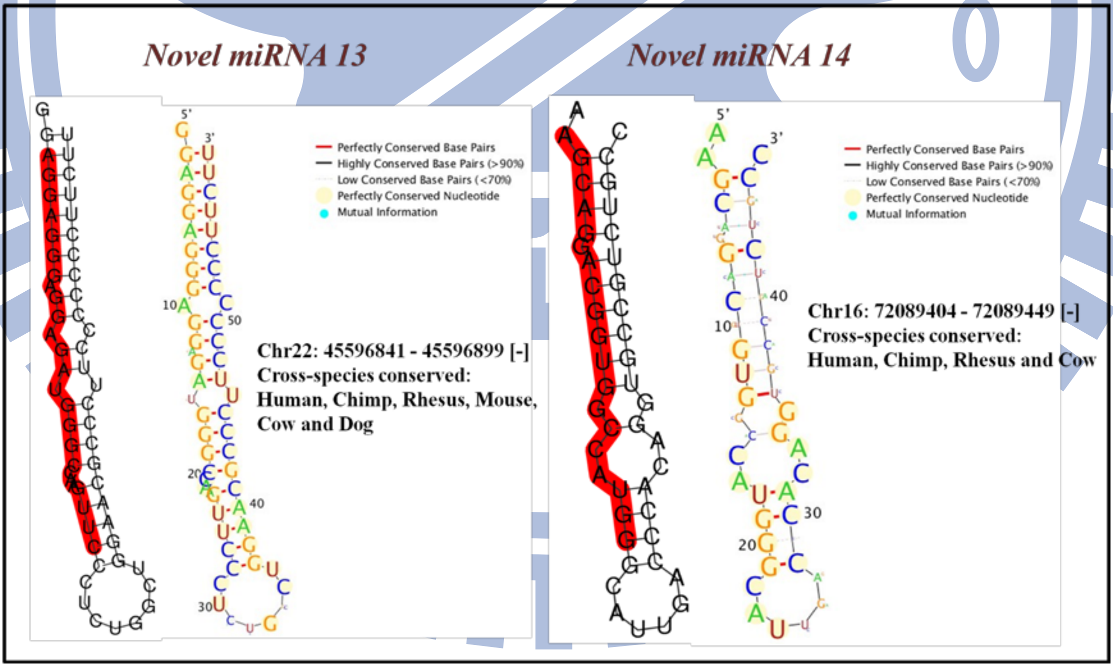

Novel miRNA 13
Novel miRNA 14
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr22: 45596841 - 45596899 [-]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Cow and Dog
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr16: 72089404 - 72089449 [-]
Cross-species conserved:
Human, Chimp, Rhesus and Cow

Figure 17. The secondary structure and cross-species conservation of novel miRNA 13 and 14 (case study 1)

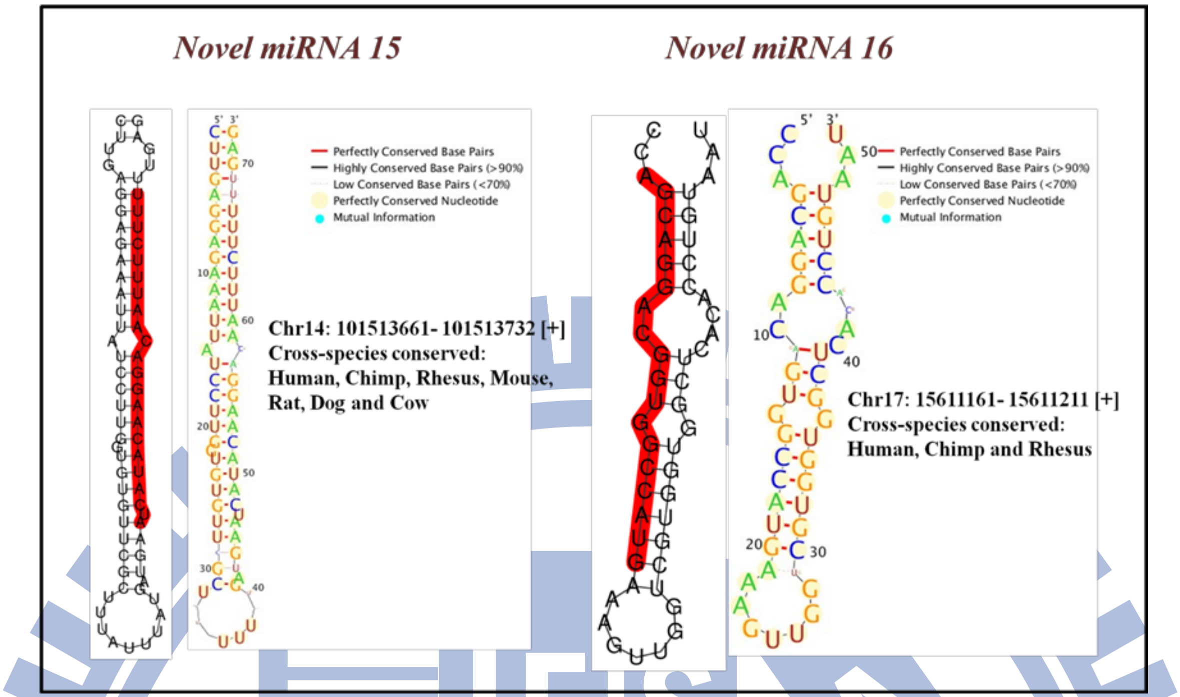

Novel miRNA 15
Novel miRNA 16
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr14: 101513661- 101513732 [+]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Rat, Dog and Cow
Chr17: 15611161- 15611211 [+]
Cross-species conserved:
Human, Chimp and Rhesus

Figure 18. The secondary structure and cross-species conservation of novel miRNA 15 and 16 (case study 1)

In profiling gene expression from microarray, 3078 differentially expressed genes are identified (fold change >=2 or <=0.5). The criteria are the same with selecting differentially expressed miRNAs. These genes are all up or down-regulated in two cancer cell lines. Then, this gene profile is used to reduce false positive rate of predicting TFs which regulate miRNAs and miRNA target sites predictions.

Table 6 lists the predicted transcription factors which regulate miRNAs. The trend of fold change of these transcription factors are the same with miRNAs. For example, hsa-miR-146a is down-regulated in cancer cells. The TF which regulates hsa-miR-146a is also down-regulated in cancer cells. hsa-miR-25 is up-regulated in cancer cells. So, the TF which regulates it is also up-regulated. Among these predicted TFs, NR3C1 id most joint TF which regulates 13 down-regulated miRNAs (total 20 down-regulated miRNAs).

Table 6. The predicted lists of transcription factor which regulate miRNAs

miRNA (fold change)	TF lists (fold change) regulate miRNAs
hsa-miR-146a (0.43)	RELA (0.44), MEIS1 (0.46), GATA3 (0.07), USF1 (0.35), REL (0.13), NFKB2 (0.11), JUN (0.19), AHR (0.01), ETS1 (0.01), NR3C1 (0.33)
hsa-miR-21 (0.45)	ZEB1 (0.23), TCF7 (0.07), RUNX3 (0.16), ETS1 (0.01), CEBPB (0.24), NR3C1 (0.33)
hsa-miR-101 (0.44)	RELA (0.44), GATA3 (0.07), EGR2 (0.30), ZEB1 (0.23), EGR3 (0.20), TCF7 (0.07), USF1 (0.35), HIF1A (0.12), AHR (0.01), TCF7L2 (0.16), SP4 (0.43), NR3C1 (0.33), STAT1 (0.33)
hsa-miR-361-3p (0.11)	MAFB (0.10), TFEC (0.12), BCL6 (0.16), TCF7 (0.07), BACH2 (0.09), MAFF (0.18), ARID5B (0.04), ETS1 (0.01), CEBPB (0.24), BACH1 (0.35), MAF (0.49), NR3C1 (0.33)
hsa-miR-27a (0.35)	GATA3 (0.07), SMAD3 (0.30)
hsa-miR-22 (0.27)	STAT5A (0.14), EGR2 (0.30), GFI1 (0.36), EGR3 (0.20), KLF12 (0.02), ARID5B (0.04), STAT4 (0.01), PAX5 (0.29), SP4 (0.43), NR3C1 (0.33), STAT1 (0.33)
hsa-miR-25 (1.54)	BRCA1 (22.29), WT1 (6.20), TFDP1 (2.10), NFYB (2.19), MYC (6.19)
hsa-miR-192 (0.65)	SMAD7 (0.18), BCL6 (0.16), JUND (0.17), FOSL2 (0.03), FOS (0.02), FOSB (0.09), JUNB (0.08), ARID5B (0.04), JUN (0.19), PAX5 (0.29), ETS1 (0.01), RORA (0.01), SMAD3 (0.30)
hsa-miR-148a (0.61)	EGR2 (0.30), EGR3 (0.20), TP53 (0.10), USF1 (0.35), KLF12 (0.02), HIF1A (0.12), ARID5B (0.04), PAX5 (0.29), MXD1 (0.16), SP4 (0.43), CEBPB (0.24)
hsa-miR-320d (0.37)	STAT5A (0.14), IRF4 (0.30), BACH2 (0.09), MAFF (0.18), TCF7L2 (0.16), SP4 (0.43), ETS1 (0.01), CEBPB (0.24), BACH1 (0.35), MAF (0.49), IRF1 (0.04), SMAD3 (0.30), MAFB (0.10), ZEB1 (0.23), IRF9 (0.15), TCF7 (0.07), USF1 (0.35), STAT4 (0.01), PAX5 (0.29), NR3C1 (0.33), STAT1 (0.33)
hsa-miR-26b (0.61)	STAT5A (0.14), RELA (0.44), EGR3 (0.20), REL (0.13), AHR (0.01), ETS1 (0.01), EGR2 (0.30), STAT4 (0.01), PAX5 (0.29), STAT1 (0.33)
hsa-miR-27b (0.41)	GCM1 (0.34), ARID5B (0.04), AHR (0.01), ETS1 (0.01), CEBPB (0.24), SMAD3 (0.30), ZEB1 (0.23), JUN (0.19), NR3C1 (0.33)
hsa-miR-181d (0.50)	EGR3 (0.20), SP4 (0.43), ETS1 (0.01), SMAD3 (0.30), EGR2 (0.30), USF1 (0.35), HIF1A (0.12), NR3C1 (0.33)
hsa-miR-152 (0.32)	EGR3 (0.20), ETS1 (0.01), EGR2 (0.30), FOS (0.02), JUN (0.19), PAX5 (0.29), NR3C1 (0.33)
hsa-miR-34c-5p (0.44)	EGR3 (0.20), AHR (0.01), SP4 (0.43), ETS1 (0.01), EGR2 (0.30), PAX5 (0.29)
hsa-miR-106b (0.48)	GFI1 (0.36), ARID5B (0.04), AHR (0.01), SP4 (0.43), ETS1 (0.01), CEBPB (0.24), USF1 (0.35), HIF1A (0.12), MXD1 (0.16), NR3C1 (0.33)
hsa-miR-151-3p (0.61)	STAT5A (0.14), EGR3 (0.20), ARID5B (0.04), AHR (0.01), SP4 (0.43), SMAD3 (0.30), EGR2 (0.30), USF1 (0.35), HIF1A (0.12), STAT4 (0.01), PAX5 (0.29), MXD1 (0.16), STAT1 (0.33)
hsa-miR-24 (0.64)	GATA3 (0.07), GCM1 (0.34), ARID5B (0.04), AHR (0.01), ETS1 (0.01), CEBPB (0.24), SMAD3 (0.30), ZEB1 (0.23), JUN (0.19), NR3C1 (0.33)
hsa-miR-26a (0.42)	STAT5A (0.14), EGR3 (0.20), TP53 (0.10), AHR (0.01), SP4 (0.43), ETS1 (0.01), EGR2 (0.30), HIF1A (0.12), STAT4 (0.01), PAX5 (0.29), NR3C1 (0.33), STAT1 (0.33)
hsa-miR-16 (0.48)	STAT5A (0.14), GFI1 (0.36), EGR3 (0.20), KLF12 (0.02), AHR (0.01), SP4 (0.43), ETS1 (0.01), IRF1 (0.04), EGR2 (0.30), USF1 (0.35), HIF1A (0.12), STAT4 (0.01), PAX5 (0.29), NR3C1 (0.33), STAT1 (0.33)
hsa-miR-1 (0.18)	EGR3 (0.20), TP53 (0.10), KLF12 (0.02), AHR (0.01), SP4 (0.43), CEBPB (0.24), IRF1 (0.04), SMAD3 (0.30), EGR2 (0.30), PAX5 (0.29), NR3C1 (0.33)

In miRNA target site predictions, the target sites of 21 differentially expressed miRNAs are identified by using miRanda and TargetScans to predict first. Then, gene expression profile is combined. If the trend of fold change between miRNAs and their target genes are the same, these genes are removed. For example, if the miRNA is down-regulated, its target gene should be up-regulated. 2864 MTIs (miRNA-target interactions) are identified from 1088 genes based on this rule. These MTIs and target genes are used to do GO and pathway enrichment analysis.

1088 predicted miRNA targets are used to do GO enrichment analysis (p-value<0.0001) by using the web-based tool (DAVID Bioinformatics Resources: http://david.abcc.ncifcrf.gov/home.jsp). Table 7 lists the analysis results of these genes. For example, 106 genes involved in DNA metabolic process, 133 genes involved in cell cycle, 57 genes involved in DNA replication, 53 genes involved in mitosis and 56 genes involved in DNA repair. These gene functions have high association with carcinogensis.

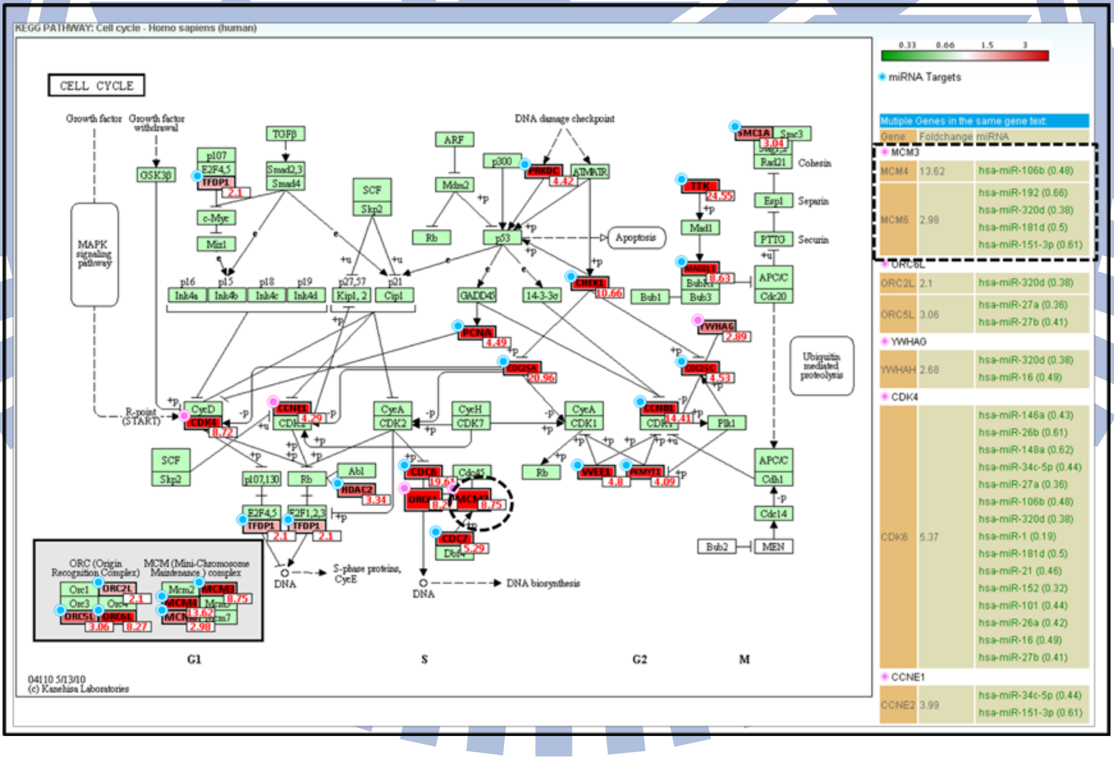
Table 7. The function of miRNA target genes by GO enrichment analysis

GO terms	Number of involved genes	p-value
GO:0006259~DNA metabolic process	106	1.03E-27
GO:0007049~cell cycle	133	6.75E-26
GO:0006260~DNA replication	57	8.12E-23
GO:0007067~mitosis	53	1.04E-16
GO:0000280~nuclear division	53	1.04E-16
GO:0006281~DNA repair	56	1.52E-13
GO:0051301~cell division	56	7.71E-13
GO:0006310~DNA recombination	25	4.41E-08
GO:0034660~ncRNA metabolic process	38	2.51E-07
GO:0006412~translation	44	9.94E-06
GO:0065003~macromolecular complex assembly	73	1.10E-05
GO:0022613~ribonucleoprotein complex biogenesis	29	1.37E-05
GO:0006886~intracellular protein transport	47	2.04E-05
GO:0032259~methylation	17	2.35E-05
GO:0007005~mitochondrion organization	24	2.58E-05
GO:0006396~RNA processing	61	4.14E-05
GO:0046907~intracellular transport	70	4.50E-05

2864 MTIs are applied for pathway enrichment analysis. Table 8 shows 17 pathways which are identified by using phyper function in R package (P-value <=0.05). For example, there are 12 miRNA target genes involved in the pathway “DNA replication”. The total number of gene in this pathway is 36. 27 miRNA target genes involve in the pathway “cell cycle”. There are total 173 genes in this pathway. Comparing the results between GO and pathway enrichment analysis, some biological processes are both identified such as DNA replication and Cell cycle. In general, the analysis result of pathway is more specific than GO enrichment analysis. For example, the pathway “Mismatch repair” is the subclass of DNA repair which is identified from GO analysis. Moreover, the pathway can offer more detailed relationship of gene and gene interaction. Figure 19 shows the miRNA target genes involve in cell cycle. The involved target genes are colored according to their fold change (red: up-regulated, green: down-regulated). In each gene text, if there are multiple genes in the same text and some of these genes are miRNA targets, pink ball is marked (blue ball means only one miRNA target gene exist in gene text). For example, the gene text “MCM” contains MCM family. Among them, MCM3, MCM4 and MCM6 are miRNA target genes. Therefore, pink ball is marked at the gene text “MCM”. MCM3 is put in the gene text. The fold changes and MTIs of MCM4 and MCM6 are put in the right side table of pathway figure. Another multiple miRNA genes in the same gene text are also marked such as YWHAG, ORC6L, CDK4 and CCNE1. Through figure 19, how these miRNA target genes lead to great influence in cell cycle can be understood more clearly.

Table 8. The pathway lists of miRNA target genes by pathway enrichment analysis

Pathway name	P-value	# of matched genes	# of genes in pathway
DNA replication	1.66E-06	12	36
Cell cycle	1.92E-05	27	173
RNA polymerase	6.55E-05	9	29
Aminoacyl-tRNA biosynthesis	4.50E-04	10	44
RNA degradation	8.09E-04	12	64
Homologous recombination	1.95E-03	8	36
Valine, leucine and isoleucine biosynthesis	2.10E-03	5	15
Alanine, aspartate and glutamate metabolism	5.83E-03	7	34
Mismatch repair	1.03E-02	6	29
Non-homologous end-joining	1.42E-02	4	15
Synthesis and degradation of ketone bodies	1.77E-02	3	9
Steroid biosynthesis	1.92E-02	6	33
p53 signaling pathway	2.44E-02	10	75
One carbon pool by folate	2.53E-02	5	26
Butanoate metabolism	2.85E-02	6	36
Pyruvate metabolism	2.95E-02	7	46
Terpenoid backbone biosynthesis	3.27E-02	4	19



Mutiple Genes in the same gene text

Gene	Foldchange	miRNA
MCM3		
MCM4	13.62	hsa-miR-106b (0.48)
MCM6	2.98	hsa-miR-192 (0.66) hsa-miR-320d (0.38) hsa-miR-181d (0.5) hsa-miR-151-3p (0.61)
ORC6L		
ORC2L	2.1	hsa-miR-320d (0.38)
ORC5L	3.06	hsa-miR-27a (0.36) hsa-miR-27b (0.41)
YWHAG		
YWHAH	2.68	hsa-miR-320d (0.38) hsa-miR-16 (0.49)
CDK4		
CDK6	5.37	hsa-miR-146a (0.43) hsa-miR-26b (0.61) hsa-miR-148a (0.62) hsa-miR-34c-5p (0.44) hsa-miR-27a (0.36) hsa-miR-106b (0.48) hsa-miR-320d (0.38) hsa-miR-1 (0.19) hsa-miR-181d (0.5) hsa-miR-21 (0.46) hsa-miR-152 (0.32) hsa-miR-101 (0.44) hsa-miR-26a (0.42) hsa-miR-16 (0.49) hsa-miR-27b (0.41)
CCNE1		
CCNE2	3.99	hsa-miR-34c-5p (0.44) hsa-miR-151-3p (0.61)

Figure 19. miRNA target genes involve in cell cycle

In case study 1, the full analysis pipeline and results are demonstrated by using public domain available data. The results contain miRNA expression profile, novel miRNAs discovery, TF list of miRNAs identification, miRNA target genes prediction, GO enrichment analysis of miRNA target genes and pathway enrichment analysis of MTIs.

Case Study 2

The data for case study 2 are from my cooperator, Dr. John Shyy, who is from Biomedical Sciences, University of California Riverside. The small RNA NGS sequencing data are under normoxia and hypoxia in human vascular endothelial cells (HUVECs). Hypoxic stress is an important factor which activates various physiological or pathophysiological responses in whole kind of cells. After processing sequencing data, 35 known differentially expressed miRNAs (fold change >=2 or <=0.5) are identified under hypoxia (Table 9). Among them, hsa-let-7 family and hsa-miR-103/107 are up-regulated. In addition, four novel miRNA candidates are found (Figure 20~21). These four novel miRNAs are conserved in human, chimp, rhesus, mouse, rat, cow and dog. HIF1A, hyroxia inducible factor 1α, is a key transcription factor induced by oxygen deprivation. It regulates hundreds of proteins involved in angiogenesis, erythropiesis, cell cycle and metastasis. To investigate the relationship between HIF1A and differentially expressed miRNAs, the binding profile of HIF1A from TRANSFAC is used to scan the promoter region of miRNAs. Transcription factor binding sites of HIFIA are identified in the promoter region of hsa-let-7 family and miR-103/107 (Figure 22). These prediction results are validated by quantifying the miRNA expression by Taqman qRCR under infecting Ad-HIF1A for 72 hr (data not shown). After verifying the relationship between HIF1A and miRNAs, the target genes of these miRNAs are predicted. 169 and 148 target genes are predicted for hsa-let-7 family and miR-103/107, respectively. Among these target genes, six genes are both targeted by hsa-let-7 family and miR-103/107 such as RPS6KB2, TARBP2, RGPD5, RGPD6, RGPD8 and EIF2C1 (Ago1). AGO1 is an important protein which anchors the miRNA-induced silencing complex (miRISC). The target sites in 3'-UTR of AGO1 are conserved in 15 species (Figure 23). Then, these two target sites are experimentally validated by Luciferase reporter assay (data not shown).

Table 9. The lists of differentially expressed miRNAs in hypoxia HUVECs

	Normalizated miRNA expression		**Fold change**
miRNA name	**normoxia**	**hypoxia**	**hypoxia/normoxia**
hsa-miR-30a*	603.94	3977.29	6.59
hsa-miR-107	5452.20	31783.39	5.83
hsa-miR-181b	423.59	2436.81	5.75
hsa-let-7e	1212.07	6628.31	5.47
hsa-let-7a	4944.72	24709.95	5.00
hsa-miR-423-5p	1056.89	5160.76	4.88
hsa-miR-886-5p	658.46	2995.88	4.55
hsa-let-7b	2461.88	11163.15	4.53
hsa-miR-320a	6899.13	30982.76	4.49
hsa-miR-26a	1103.02	4726.26	4.28
hsa-let-7g	3422.30	14490.22	4.23
hsa-miR-31	1665.02	7023.31	4.22
hsa-let-7f	15815.57	52860.90	3.34
hsa-miR-103	46788.24	148022.67	3.16
hsa-miR-16	549.41	1636.19	2.98
hsa-let-7c	4122.70	10830.45	2.63
hsa-miR-432	436.18	1109.02	2.54
hsa-miR-93	1048.50	2573.54	2.45
hsa-miR-26b	1119.80	2710.27	2.42
hsa-let-7d	4043.01	9138.05	2.26
hsa-miR-221	11709.64	24904.40	2.13
hsa-miR-29c	759.11	376.76	0.50
hsa-miR-106b	998.17	487.67	0.49
hsa-miR-7	1258.20	572.74	0.46
hsa-miR-101	11915.15	5422.06	0.46
hsa-miR-379	4198.19	1861.03	0.44
hsa-miR-340	1212.07	528.69	0.44
hsa-miR-216a	675.23	265.86	0.39
hsa-miR-21	562171.82	206790.26	0.37
hsa-miR-128	662.65	206.61	0.31
hsa-miR-411	2264.76	598.57	0.26
hsa-miR-30c	977.20	212.69	0.22
hsa-miR-23b	1342.08	278.02	0.21
hsa-miR-23a	10916.98	2193.74	0.20
hsa-miR-126*	6349.71	1157.64	0.18

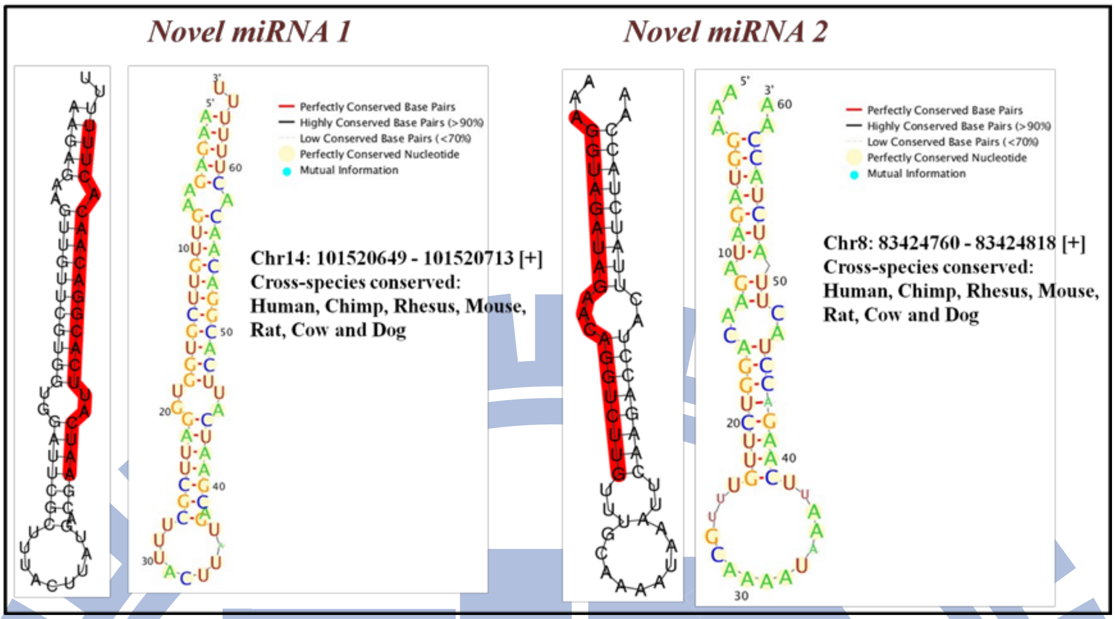

Novel miRNA 1
Novel miRNA 2
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr14: 101520649 - 101520713 [+]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Rat, Cow and Dog
Chr8: 83424760 - 83424818 [+]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Rat, Cow and Dog

Figure 20. The secondary structure and cross-species conservation of novel miRNA 1 and 2 (case study 2)

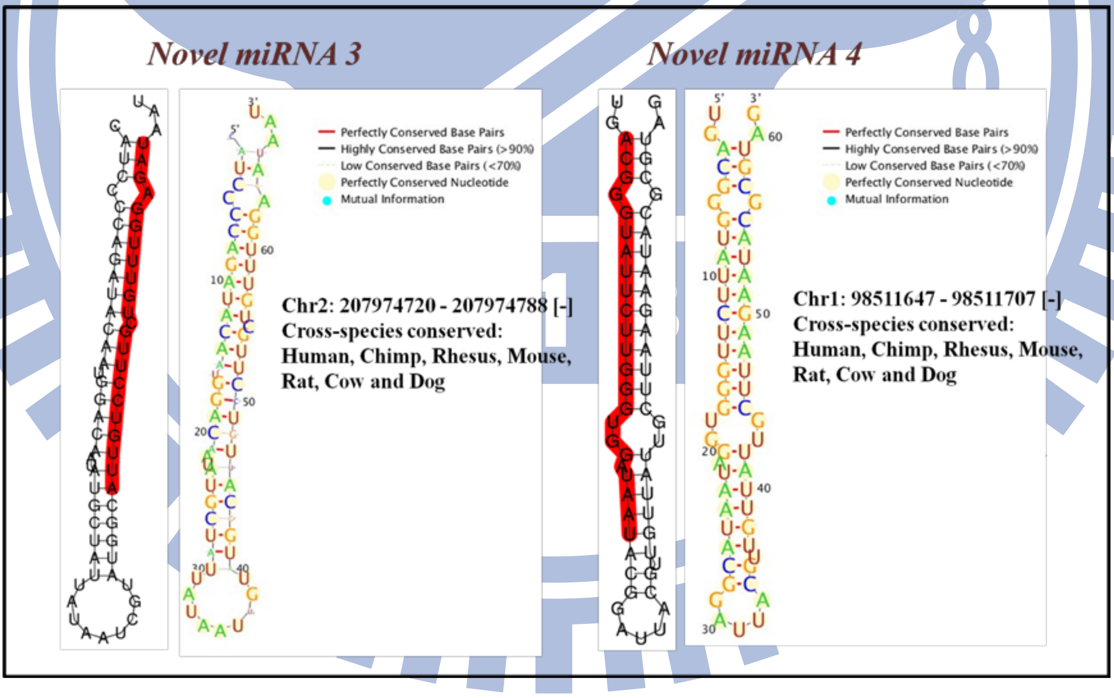

Novel miRNA 3
Novel miRNA 4
Perfectly Conserved Base Pairs
Highly Conserved Base Pairs (>90%)
Low Conserved Base Pairs (<70%)
Perfectly Conserved Nucleotide
Mutual Information
Chr2: 207974720 - 207974788 [-]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Rat, Cow and Dog
Chr1: 98511647 - 98511707 [-]
Cross-species conserved:
Human, Chimp, Rhesus, Mouse,
Rat, Cow and Dog

Figure 21. The secondary structure and cross-species conservation of novel miRNA 3 and 4 (case study 2)

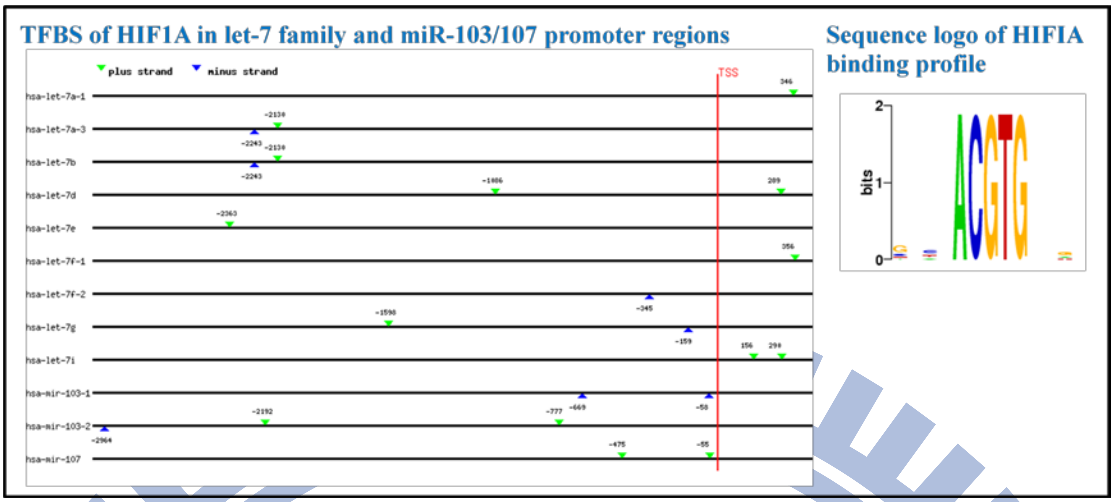


Figure 22. HIFIA binding sites in the promoter regions of let-7 family and miR-103/107

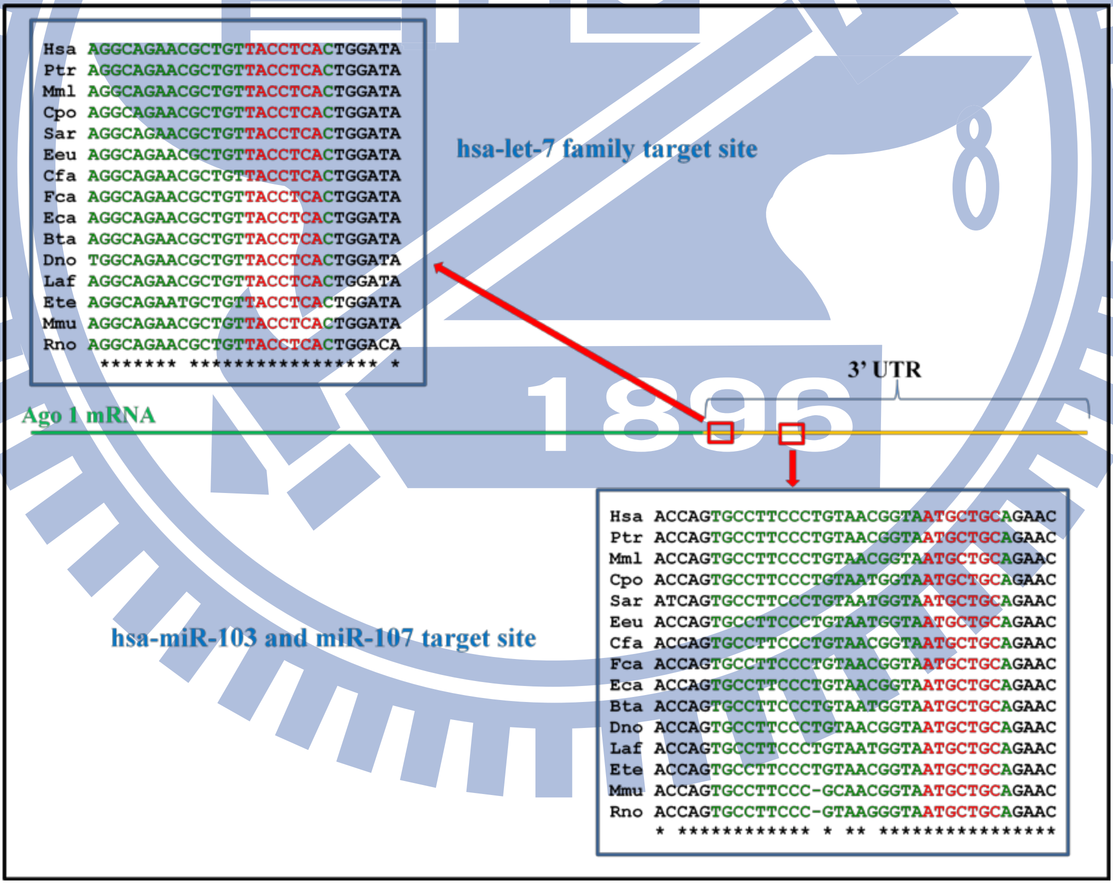

Hsa AGGCAGAACGCTGTTACCTCACTGGATA
Ptr AGGCAGAACGCTGTTACCTCACTGGATA
Mml AGGCAGAACGCTGTTACCTCACTGGATA
Cpo AGGCAGAACGCTGTTACCTCACTGGATA
Sar AGGCAGAACGCTGTTACCTCACTGGATA
Eeu AGGCAGAACGCTGTTACCTCACTGGATA
Cfa AGGCAGAACGCTGTTACCTCACTGGATA
Fca AGGCAGAACGCTGTTACCTCACTGGATA
Eca AGGCAGAACGCTGTTACCTCACTGGATA
Bta AGGCAGAACGCTGTTACCTCACTGGATA
Dno TGGCAGAACGCTGTTACCTCACTGGATA
Laf AGGCAGAACGCTGTTACCTCACTGGATA
Ete AGGCAGAATGCTGTTACCTCACTGGATA
Mmu AGGCAGAACGCTGTTACCTCACTGGATA
Rno AGGCAGAACGCTGTTACCTCACTGGACA
******* ***************** *
hsa-let-7 family target site
3' UTR
Ago 1 mRNA
hsa-miR-103 and miR-107 target site
Hsa ACCAGTGCCTTCCCTGTAACGGTAATGCTGCAGAAC
Ptr ACCAGTGCCTTCCCTGTAACGGTAATGCTGCAGAAC
Mml ACCAGTGCCTTCCCTGTAACGGTAATGCTGCAGAAC
Cpo ACCAGTGCCTTCCCTGTAATGGTAATGCTGCAGAAC
Sar ATCAGTGCCTTCCCTGTAATGGTAATGCTGCAGAAC
Eeu ACCAGTGCCTTCCCTGTAATGGTAATGCTGCAGAAC
Cfa ACCAGTGCCTTCCCTGTAACGGTAATGCTGCAGAAC
Fca ACCAGTGCCTTCCCTGTAACGGTAATGCTGCAGAAC
Eca ACCAGTGCCTTCCCTGTAACGGTAATGCTGCAGAAC
Bta ACCAGTGCCTTCCCTGTAATGGTAATGCTGCAGAAC
Dno ACCAGTGCCTTCCCTGTAACGGTAATGCTGCAGAAC
Laf ACCAGTGCCTTCCCTGTAATGGTAATGCTGCAGAAC
Ete ACCAGTGCCTTCCCTGTAACGGTAATGCTGCAGAAC
Mmu ACCAGTGCCTTCCC-GCAACGGTAATGCTGCAGAAC
Rno ACCAGTGCCTTCCC-GTAAGGGTAATGCTGCAGAAC
* ************ * ** *****************

Figure 23. The target sites of hsa-let-7 family and hsa-miR-103/107 in the 3' UTR of Ago1. These target sites are conserved in 15 species (red part is seed region).

According to *in silico* bioinformatics analysis and *in vitro* validation results, the model "miRNA-mediated transcriptional de-suppression" is proposed (Figure 24). In this model, HIF1A is induced under hypoxia. Then, HIF1A induces the expression of hsa-let-7 family and miR-103/107. These miRNAs down-regulate the expression level of Ago1 by targeting the 3'-UTR of it. The quantities of miRISC formation are reduced due to the low expression level of Ago1. So, the translational repression by miRNA targeting is reduced. The genes which are targeted by miRNAs are up-regulated. To verify this model, VEGF (vascular endothelial growth factor), induced under hypoxia, is validated about "transcriptional de-suppression" *in virto* and *in vivo* (data not shown).

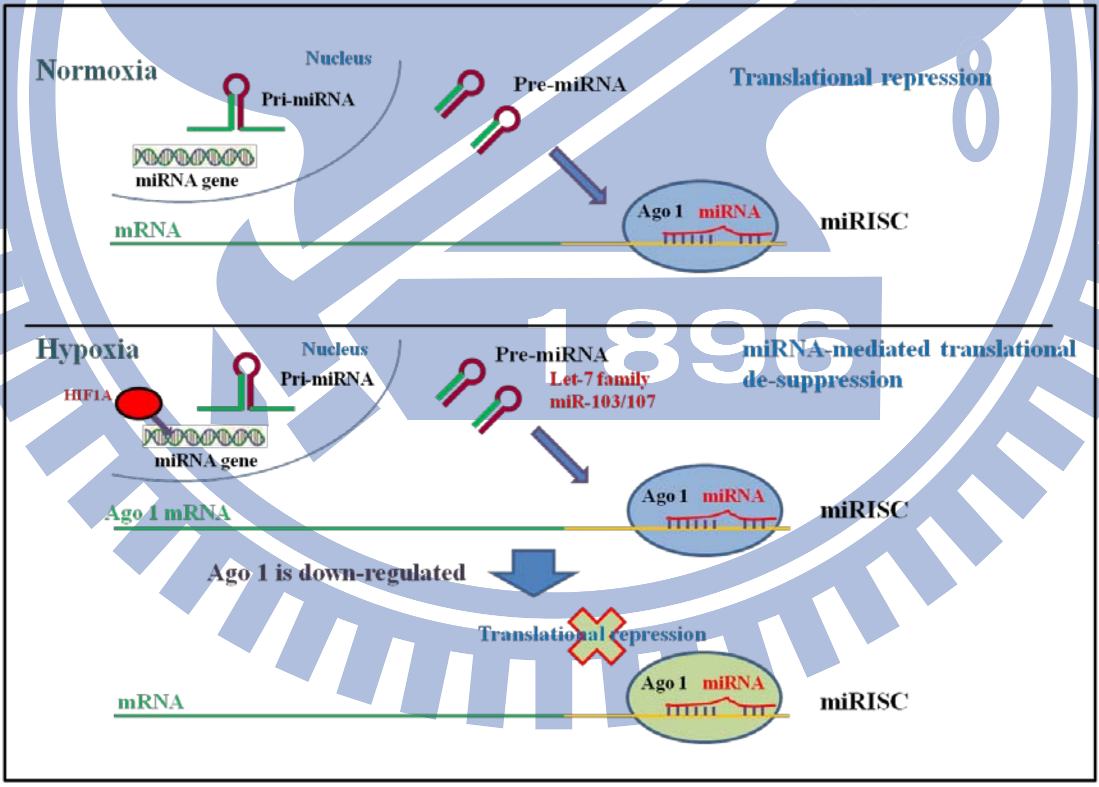


Figure 24. The model of translational suppression under normoxia and de-suppression under hypoxia

Case Study 3

The small RNA NGS sequencing data are from my cooperator, Dr. Hailing Jin, who is from Department of Plant Pathology and Microbiology, University of California Riverside [173]. There are four immunoprecipitation sequencing data sets from *Arabidopsis* leaves. The selected proteins are Ago1 and Ago2. These four sets are AGO1-IP-mock, AGO1-IP-*Pst (Pseudomonas syringae pv. tomato)*, AGO2-IP-mock and AGO2-IP-*pst*. Excluding these four sets, there are two sequencing data about mutant Ago2 (ago2-mock and ago2-*pst*). Argonaute (AGO) proteins are important protein family in plants. They are RNAi effectors that bind miRNAs or siRNAs and mediate gene silencing by targeting the mRNAs. In plants, there are various immune responses including pathogen-associated molecular pattern-triggered immunity (PTI) and bacterial effector-triggered immunity (ETI). In plants, AGO1 is primarily associated with miRNAs and regulates PTI by several stress-related miRNAs. In figure 25, the mRNA level of AGO2 is highly induced by virulent *Pst* (EV) and avirulent *Pst* (*avrRpt2*) at 6 and 14 hours post inoculation (hpi). Therefore, AGO2 is the RNA silencing effector in both ETI and PTI by bounding miRNAs or siRNAs. To investigate the function of AGO2 in antibacterial immune responses, AGO1- and AGO2-IP SBS sequencing data are generated. AGO1-IP libraries are used as control. ago2 mutant sequencing data by the same treatment are also constructed.

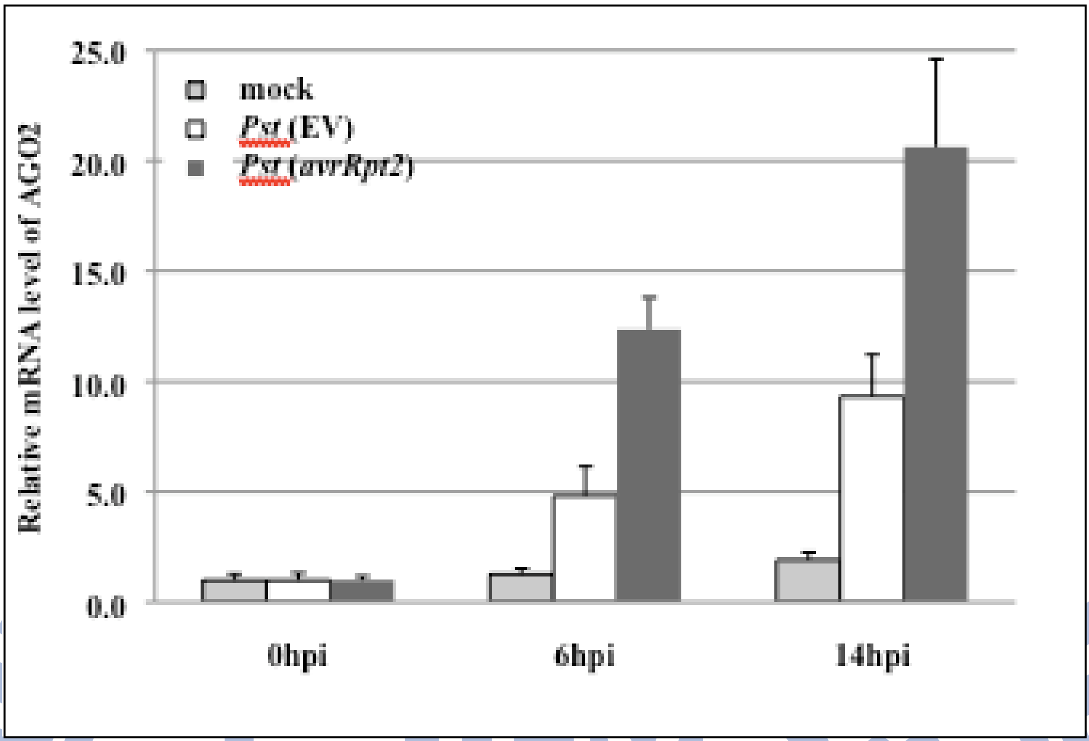


Figure 25. The mRNA expression level of AGO2 in mock, virulent *Pst* (EV) and avirulent *Pst* (*avrRpt2*)

After processing the sequencing data and profiling the expression level of miRNAs and various kind siRNAs, an interesting thing is found. The expression level of various miRNA and miRNA* are opposite in AGO1- and AGO2-IP sets. For example, ath-miR165a is abundant in AGO1-IP but is lowly expressed in AGO2-IP. ath-miR165a* is abundant in AGO2-IP but is lowly expressed in AGO1-IP. In table 10, there are five miRNA and miRNA* pairs which have great different expression level between AGO1- and AGO2-IP sets. In previous studies, only one strand of miRNA is expressed in many experimental conditions and another strand of miRNA is usually considered as non-function.

Table 10. The lists of differentially expressed miRNAs (comparing AGO1- and AGO2-IP)

	AGO1-IP		AGO2-IP		ago2	
miRNA name	mock	*Pst (avrRpt2)*	mock	*Pst (avrRpt2)*	mock	*Pst (avrRpt2)*
ath-MIR165a	30890.43	23823.16	0.29	6.34	24.30	54.25
ath-MIR165a*	0.98	0.93	7943.39	15529.25	23.54	1050.87
ath-MIR393b	358.14	156.56	0.44	0.52	2.85	8.60
ath-MIR393b*	0.33	24.40	2621.86	4546.56	5.51	475.12
ath-MIR396a	1071.49	1258.68	18.68	13.07	33.79	632.99
ath-MIR396a*	9457.78	9165.25	1016.89	2158.92	555.69	1900.09
ath-MIR396b	4121.88	6172.96	26.32	28.40	97.58	1548.06
ath-MIR396b*	188.20	214.62	1196.89	2864.61	26.01	255.14
ath-MIR472	246.91	61.45	0.88	2.01	2.09	16.09
ath-MIR472*	28.05	36.44	508.67	1273.69	0.00	63.60

To investigate the function of the miRNA*, the target sites of these five miRNA* are predicted by the modified plant target prediction guideline (described in the “Materials and Methods” section). In table 11, there are 41 predicted genes which are targeted by these five miRNA* (16 target genes for miR165a*, 13 target genes for miR393b*, 6 target genes for miR396a*, 3 target genes for miR396b* and 3 target genes for miR472*). The function of miR393b has been reported that it contributes to PTI by silencing auxin receptors TIR1, AFB2, and AFB3. To test whether miR393b* also involves in PTI, AT5G50440 encodes MEMB12, an SDS-resistant soluble N-ethylmaleimide sensitive factor attachment protein receptor (SNARE) that is mainly localized in *cis*-Golgi cisternae. The relationship between MEMB12 and miR393* is validated by detecting the protein level of wild-type MEMB12 (MEMB12-wt) and the miR393* target site mutated in MEMB12 (MEMB12-mu). In figure 26, the protein level of MEMB12 is reduced by miR393* targeting in MEMB12-wt but the protein level of MEMB12 is recovered by mutating miR393* target site in MEMB12-mu.

Table 11. The predicted target genes of selected miRNA*

miRNA name	Locus ID	Description	Score
ath-miR165a*	AT2G30490	C4H (CINNAMATE-4-HYDROXYLASE); trans-cinnamate 4-monooxygenase	4
ath-miR165a*	AT1G66150	TMK1 (TRANSMEMBRANE KINASE 1); transmembrane receptor protein serine/threonine kinase	4
ath-miR165a*	AT1G74790	catalytic	4
ath-miR165a*	AT2G02790	IQD29 (IQ-domain 29); calmodulin binding	4.5
ath-miR165a*	AT5G06350	binding	4.5
ath-miR165a*	AT5G52620	F-box family protein	5
ath-miR165a*	AT3G17850	protein kinase, putative	5
ath-miR165a*	AT3G19650	cyclin-related	5
ath-miR165a*	AT4G31600	UDP-glucuronic acid/UDP-N-acetylgalactosamine transporter-related	5.5
ath-miR165a*	AT3G28340	GATL10 (Galacturonosyltransferase-like 10); polygalacturonate 4-alpha-galacturonosyltransferase/ transferase, transferring hexosyl groups	5.5
ath-miR165a*	AT2G42300	basic helix-loop-helix (bHLH) family protein	5.5
ath-miR165a*	AT5G44400	FAD-binding domain-containing protein	5.5
ath-miR165a*	AT5G61780	tudor domain-containing protein / nuclease family protein	5.5
ath-miR165a*	AT5G01030	unknown protein	5.5
ath-miR165a*	AT5G54690	GAUT12 (GALACTURONOSYLTRANSFERASE 12); polygalacturonate 4-alpha-galacturonosyltransferase/ transferase, transferring glycosyl groups / transferase, transferring hexosyl groups	5.5
ath-miR165a*	AT3G48580	xyloglucan:xyloglucosyl transferase, putative / xyloglucan endotransglycosylase, putative / endo-xyloglucan transferase, putative	5.5
ath-miR393b*	AT3G63180	TKL (TIC-LIKE)	4.5
ath-miR393b*	AT3G19420	ATPEN2 (ARABIDOPSIS THALIANA PTEN 2); phosphatase/ protein tyrosine phosphatase	4.5
ath-miR393b*	AT1G61350	armadillo/beta-catenin repeat family protein	5
ath-miR393b*	AT3G48350	cysteine proteinase, putative	5
ath-miR393b*	AT4G19490	protein binding	5
ath-miR393b*	AT4G09040	RNA recognition motif (RRM)-containing protein	5
ath-miR393b*	AT2G39300	unknown protein	5
ath-miR393b*	AT5G58660	oxidoreductase, 2OG-Fe(II) oxygenase family protein	5
ath-miR393b*	AT2G19640	ASHR2 (ASH1-RELATED PROTEIN 2)	5.5
ath-miR393b*	AT2G25140	CLPB4 (CASEIN LYTIC PROTEINASE B4); ATP binding / ATPase/ nucleoside-triphosphatase/ nucleotide binding / protein binding	5.5
ath-miR393b*	AT5G05900	UDP-glucoronosyl/UDP-glucosyl transferase family protein	5.5
ath-miR393b*	AT5G50440	MEMB12 (MEMBRIN 12); SNAP receptor	5.5
ath-miR393b*	AT3G09530	ATEXO70H3 (exocyst subunit EXO70 family protein H3); protein binding	5.5
ath-miR396a*	AT2G19590	ACO1 (ACC OXIDASE 1); 1-aminocyclopropane-1-carboxylate oxidase	4.5
ath-miR396a*	AT5G15610	proteasome family protein	5
ath-miR396a*	AT2G18710	SCY1 (SecY Homolog 1); P-P-bond-hydrolysis-driven protein transmembrane transporter	5
ath-miR396a*	AT1G51340	MATE efflux family protein	5.5
ath-miR396a*	AT5G14130	peroxidase, putative	5.5
ath-miR396a*	AT3G18220	phosphatidic acid phosphatase family protein / PAP2 family protein	5.5
ath-miR396b*	AT1G19720	pentatricopeptide (PPR) repeat-containing protein	4.5
ath-miR396b*	AT1G09610	unknown protein	5
ath-miR396b*	AT5G60610	F-box family protein	5.5
ath-miR472*	AT4G13395	RTFL12 (ROTUNDIFOLIA LIKE 12)	5
ath-miR472*	AT1G12290	disease resistance protein (CC-NBS-LRR class), putative	5
ath-miR472*	AT2G28010	aspartyl protease family protein	5.5

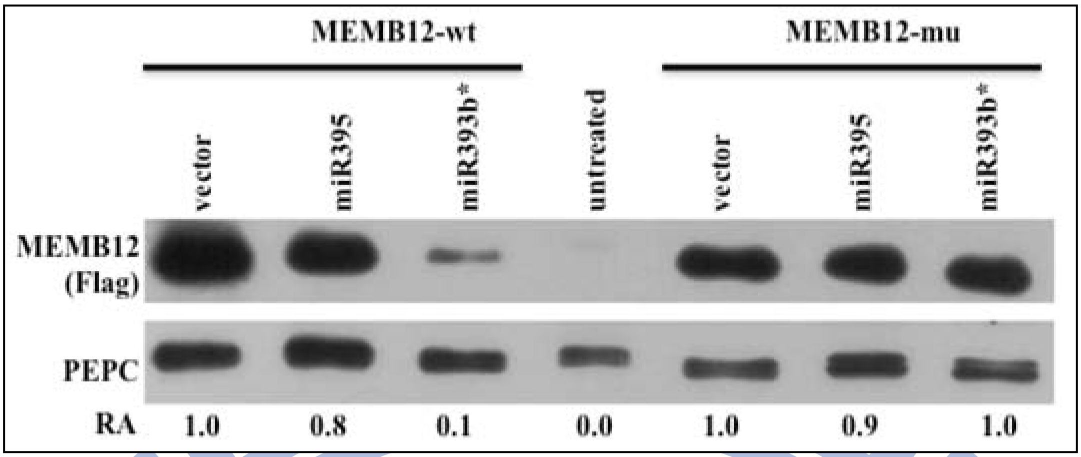
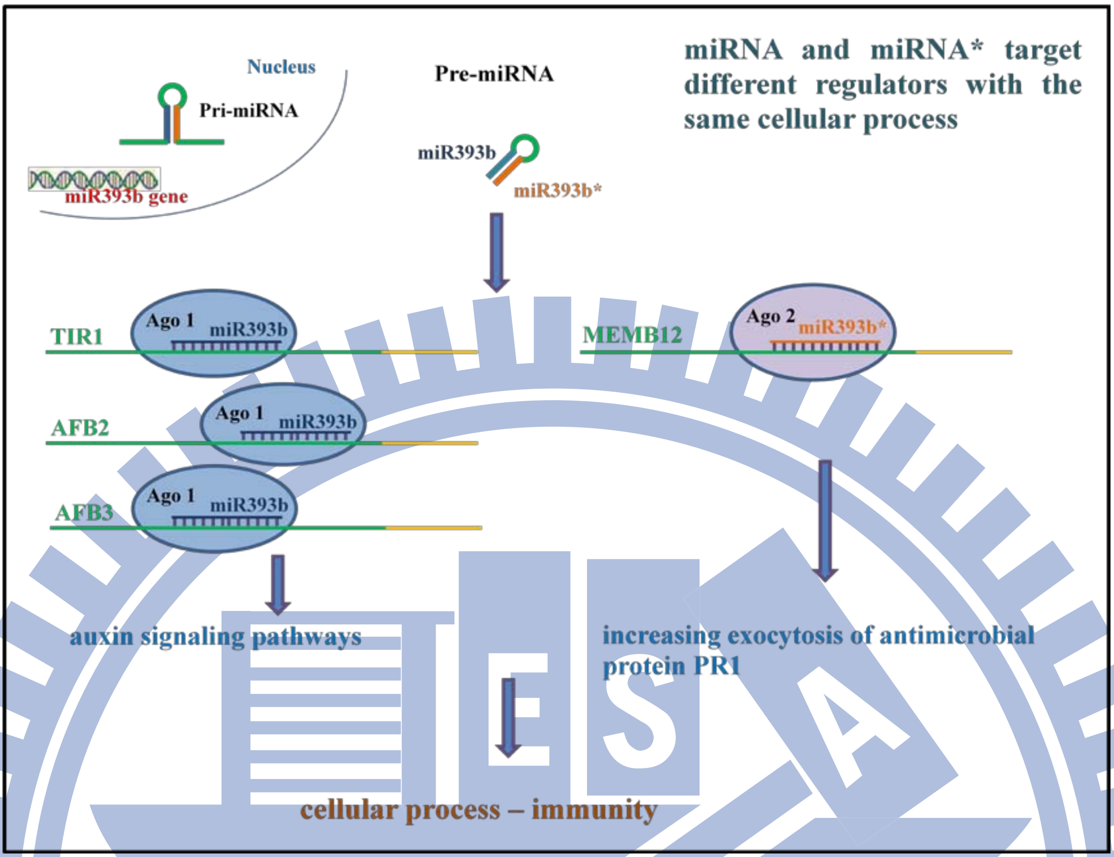


Figure 26. The protein level of MEMB12 with miR393b* (MEMB12-wt and mu)

The model of miRNA* and miRNA pair associating with different AGO proteins targeting different regulators for the same cellcular process (PTI) is proposed based on in silico bioinformatics and in virto experimental validated results. Figure 27 demonstrates this model. miR393b associated with AGO1 targets auxin receptors TIR1, AFB2, and AFB3 for regulating pattern-triggered immunity (PTI). miR393b* associated with AGO2 targets MEMB12. Reduced *MEMB12* leads to increased exocytosis of antimicrobial protein PR1. Therefore, AGO2 regulates PTI by binding miR393b* and subsequently modulating exocytosis of antimicrobial molecules.

Excluding miR393b and miR393b* pair, there are other miRNA nad miRNA* pairs identified in this study. AGO2 is induced by ETI and PTI. Therefore, the function of miRNA* associated with AGO2 may regulate a group of genes involved in various pathways which are correlated with plant immunity.



Nucleus
Pri-miRNA
miR393b gene
Pre-miRNA
miR393b
miR393b*
miRNA and miRNA* target different regulators with the same cellular process
TIR1
Ago 1
miR393b
AFB2
Ago 1
miR393b
AFB3
Ago 1
miR393b
MEMB12
Ago 2
miR393b*
auxin signaling pathways
increasing exocytosis of antimicrobial protein PR1
cellular process – immunity

Figure 27. miRNA* and miRNA pair, each of which targets different regulators within the same cellular process – immunity through two distinct RNAi effectors.

Case Study 4

The small RNA sequencing data are from my cooperator, Dr. Lin Na-Sheng, who is from Institute of Plant and Microbial Biology, ACADEMIA SINICA. In plants, viruses can induce post-transcriptional gene silencing (PTGS) through virus-specific small interfering RNAs (vsiRNAs) targeting the mRNA of host. Satellite RNAs (satRNAs) can also trigger PTGS to produce satRNA-derived siRNAs (satsiRNAs). In Dr. Lin's previous studies, two BaMV (*Bamboo mosaic virus*)-associated Satellite RNAs (satBaMVs) are identified [174]. They are BSL6 and BSF4. The similarity between BSL6 and BSF4 is very high (93%). BSL6 can greatly reduce BaMV accumulation and attenuate BaMV-induced symptoms in *N. benthamiana* and *Chenopodium quinoa*. However, BSF4 does not have this function. The key component of satBaMV determines this function is the apical hairpin stem loop (AHSL) located in the 5'-UTR of BSL6. To further investigate the relationship between BaMV and satBaMVs, total eleven sequencing sets from *Arabidposis thaliana* and *Nicotiana benthamiana* are generated [175]. There are 4 sets from inoculated leaves of *A. thaliana* (Mock, BaMV, BaMV+BSF4 and BaMV+BSF6) and 7 sets from inoculated (I) and systemic (S) leaves of *N. benthamiana* (Mock, BaMV(I), BaMV(S), BaMV+BSF4(I), BaMV+BSF4(S), BaMV+BSL6(I) and BaMV+BSL6(S)).

After processing sequencing data, the amount of vsiRNAs and satsiRNAs are counted in each library. Table 12 lists the amount of vsiRNAs and satsiRNAs in *N. benthamiana.* In inoculated and systemic leaves with BSL6, very few vsiRNAs and satsiRNAs are detected. It is corresponded to the results of previous study (low expression level of BaMV and BSL6 RNAs) [176]. In *A. thaliana*, the expression level of vsiRNAs and satsiRNAs is also very low (Table 13). The detected expression level of vsiRNAs and satsiRNAs in BSF4 is higher than in BSL6. It is because BSF4 can not reduce BaMV accumulation (the higher expression level of BaMV and BSF4 RNAs). According the statistics of length of vsiRNAs and satsiRNAs, the major length is 21 nt in *A. thaliana* (Figure 28). The major length is 21 and 22 nt in *N. benthamiana* (Figure 29).

Table 12. The amount of small RNAs in 7 libraries in *N. benthamiana*

	Mock	BaMV (I)	BaMV (S)	BaMV +BSF4 (I)	BaMV +BSF4 (S)	BaMV +BSL6 (I)	BaMV +BSL6 (S)
Total	2,489,506	3,293,585	3,411,741	5,247,031	3,567,466	3,272,499	4,853,059
BaMV		123,901 (3.7%)	596,851 (17.5%)	35,794 (0.7%)	829,730 (23.3%)	2,041 (0.1%)	95 (0.0%)
BSF4				269,101 (5.1%)	53,539 (1.5%)		
BSL6						7,441 (0.2%)	71 (0.0%)

Table 13. The amount of small RNAs in 4 libraries in *A. thaliana*

	Mock	BaMV	BaMV+BSF4	BaMV+BSL6
Total	4,676,816	3,437,925	2,037,033	2,221,999
BaMV		23,714 (0.7%)	29,880 (1.5%)	4,555 (0.2%)
BSF4			1,281(0.1%)	
BSL6				1,653 (0.1%)

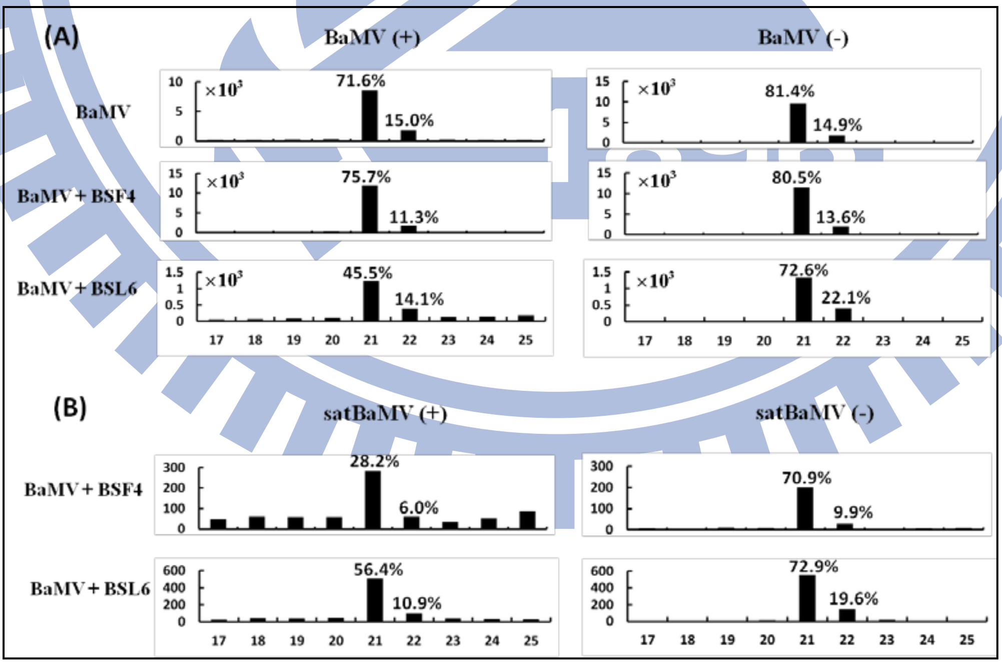


Figure 28. Length distribution of siRNAs derived from BaMV or satBaMV in *A. thaliana*

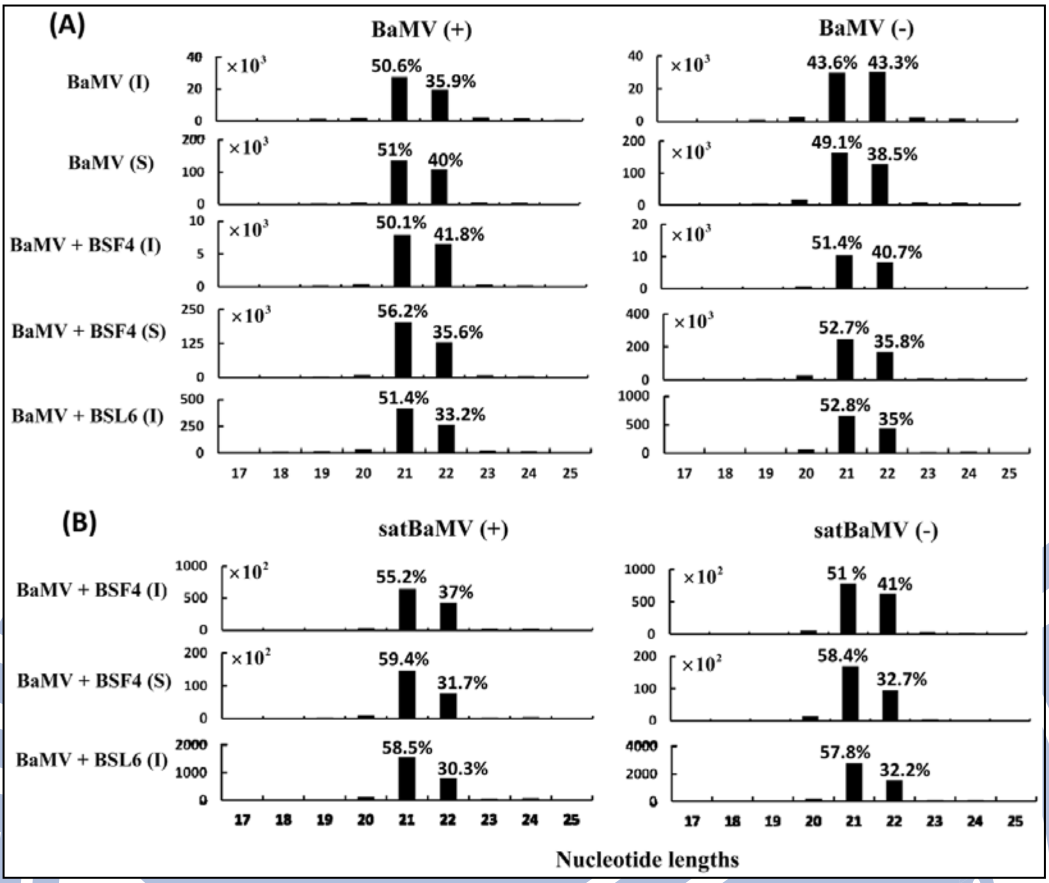


Figure 29. Length distribution of siRNAs derived from BaMV or satBaMV in *N. benthamiana*

To analyze whether any bias in length distribution of siRNAs in mock, virus-infected and satRNA-co-infected samples, the length distribution of total siRNAs are compared. Obviously, the length distribution of total siRNAs is different with vsiRNAs and satsiRNAs (Figure 30). The most abundant siRNAs is 22 nt in mock. After BaMV inoculation or co-inoculation with satBaMVs, the quantity of 22 nt siRNAs are greatly decreased and 21 and 24 nt siRNAs are increased. These 24 nt siRNAs are endogenous siRNAs which belong to host (24 nt siRNAs are very few in BaMV and satBaMV).

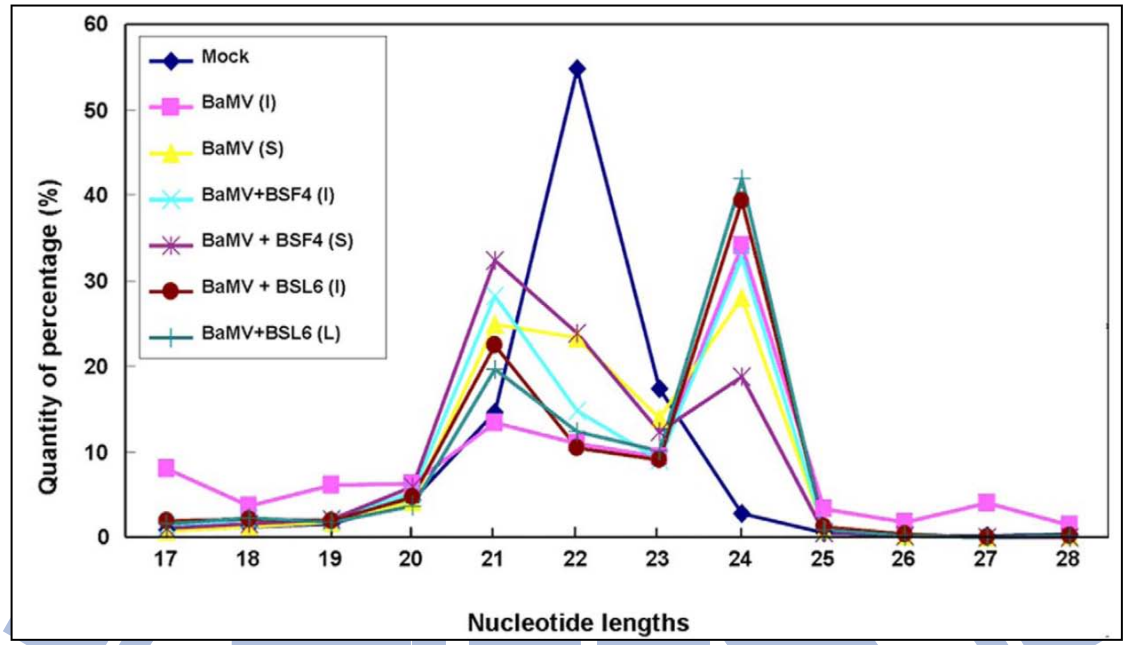


Figure 30. The length distribution of total siRNAs from *N. benthamiana*

Figure 31 shows the distribution of siRNAs in the 5'-UTR of satRNAs in *A. thaliana*. and *N. benthamiana*. The key component of the satRNA for reducing BaMV accumulation is called as AHSL region (position 54-91 in BSF4 and position 54-92 in BSL6). In AHSL region, there are two major peaks in BSF4 negative strand (position 54 and 75). There are several peaks in BSL6 negative strand (position 51-61 and position 71-81).

According to these analysis results, the low expression level of vsiRNAs and satsiRNAs of BSL6 and its correlation with the accumulation of BaMV and satBaMV suggests that BSL6 down-regulates BaMV before triggering PTGS (post-transcriptional gene silencing) to produce siRNAs. Moreover, the length distributions of siRNAs with and without BSL6 co-infection are similar (Figure 28). These results also imply that reduced BaMV accumulation by BSL6 occurs before PTGS.

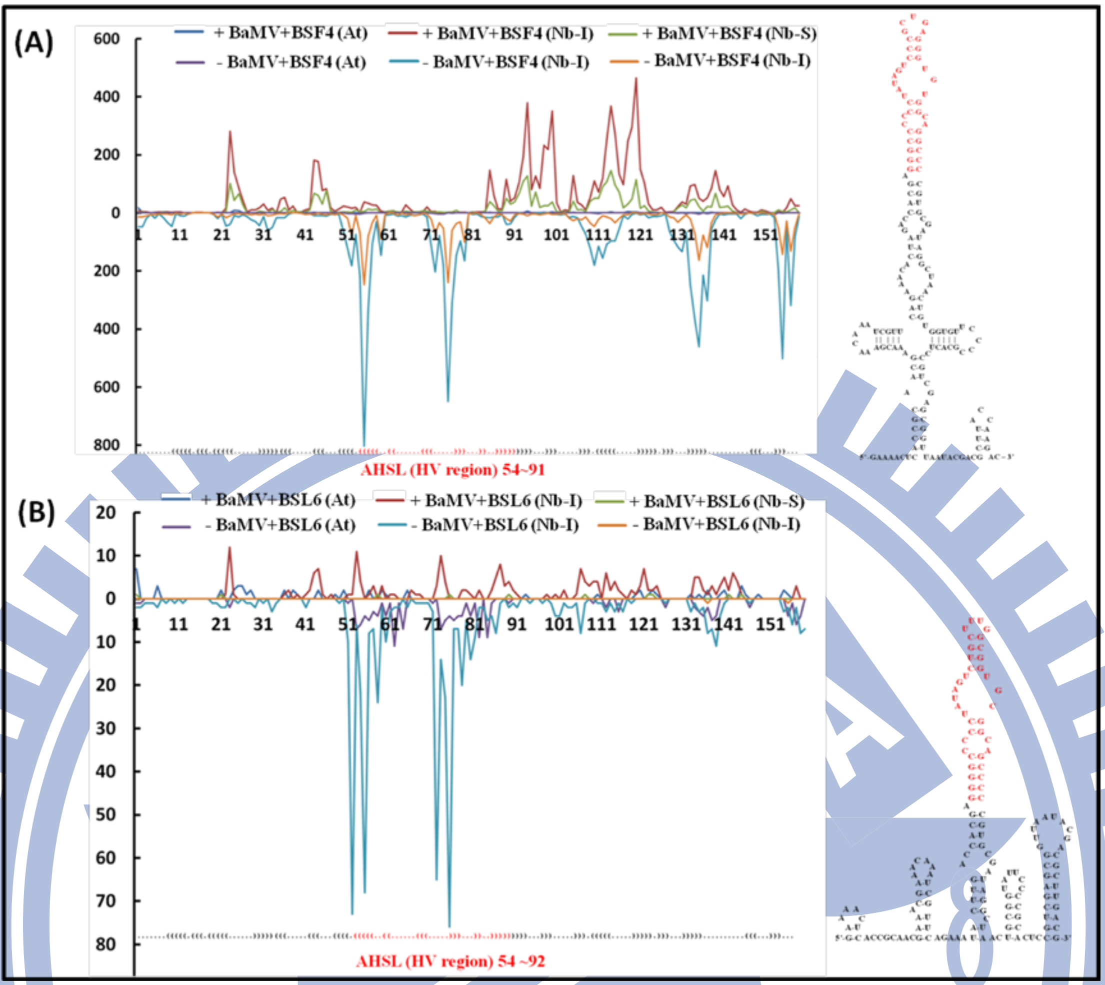


Figure 31. The siRNA distributions in 5' UTR of satBaMV in *A. thaliana*. and *N. benthamiana* (A. BSF4, B. BSL6).

6. Discussions

My work purposes the complete, systematic and comprehensive analysis pipeline for small non-coding RNA next-generation sequencing data. It can help biologists to easily process their own data and find biological significant genes or pathways after running whole processes. In results section, case study 1 clearly demonstrates that the full analysis results are generated by using public domain available data. Case studies 2-4 are the real results from my cooperators. Case study 2 is finding the miRNA regulation role under normoxia and hypoxia in human vascular endothelial cells (HUVECs). After the analysis flow and experimental validated, the model "miRNA-mediated transcriptional de-suppression" is proposed (hsa-miR-let 7 family and miR-103/107 targeting AGO1). In case study 3, the function of AGO protein in antibacterial immune responses in *Arabidopsis* is investigated. According to *in silico analysis and in virto* experimental results, the miRNA and miRNA* pair associated with different AGO proteins targeting different regulators with the same cellular process is proposed. In case study 4, the sequencing data of small RNA of virus (BaMV) and virus-associated satellite RNA (BSF4 and BSF6) in different plants are generated and analyzed. The characteristics of vsiRNAs and satsiRNAs are found (the major length of these siRNAs). And BSF6 reduced BaMV accumulation triggering PTGS to produce siRNAs is also found.

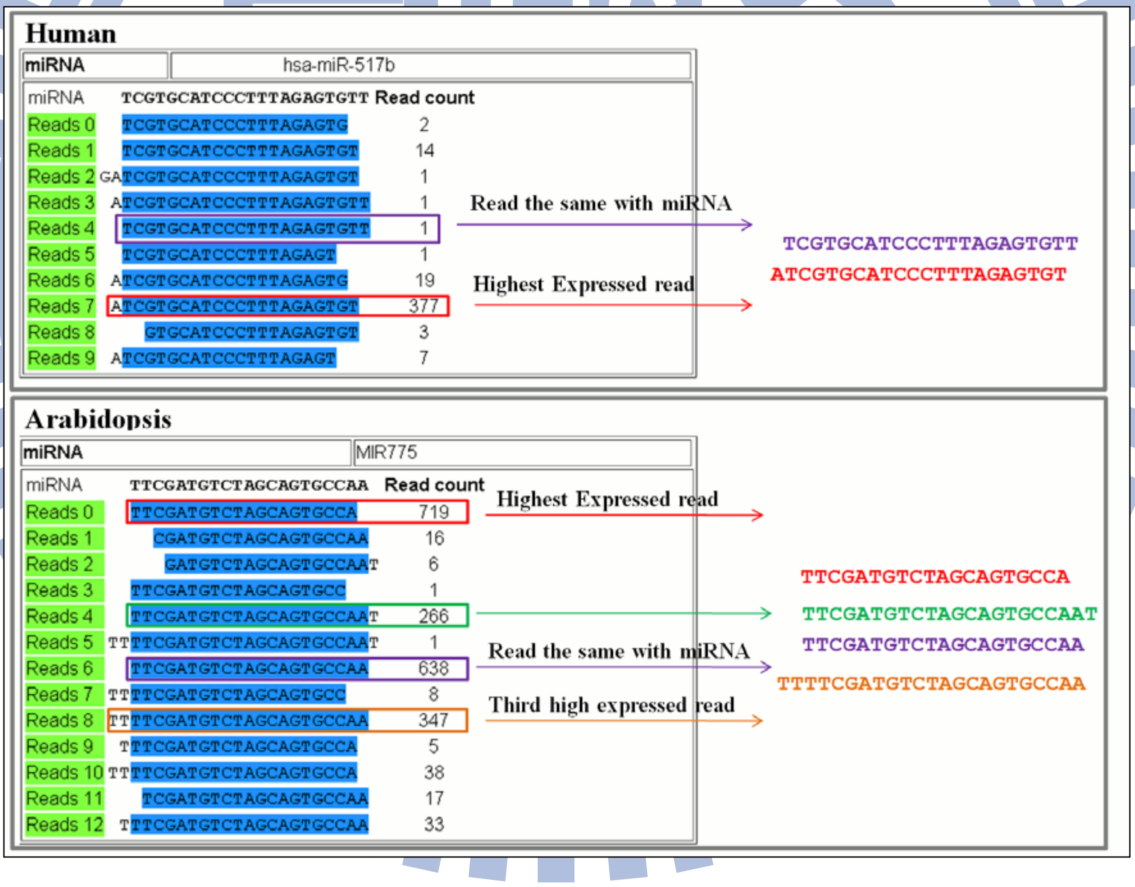
The systematic analysis flow for small non-coding RNA NGS sequencing data is powerful and useful for biologists. However, increasing the correctness of profiling NGS data and how to combine various kinds of NGS data need to be further investigated.

6.1 Increasing the correctness of profiling NGS data

Undoubtedly, next-generation sequencing technology is a powerful approach for profiling and analyzing small non-coding RNAs. However, how to monitor the expression level of small RNAs more correctly is an important issue. The major problems are how to deal with multiple mapping reads (mutlireads) and the various sequencing reads for known small non-coding RNAs (In miRNAs, they are called as isomiRs).

Multiple mapping reads represent the reads which have multiple genomic locations. For instance, in Illumina platform, the length of the read after trimming 3’ adaptor sequence is smaller than 30 nt. Then, these reads are mapped to genomic sequences. Some of them can be mapped to multiple locations. About 30% of the total mapped reads are mutireads [177]. In various previous studies, these multiple mapping reads are removed before further analysis. However, discarding mutireads affects the detected expression levels [178]. Several studies have proposed the solutions based on heuristic approach which divides mutireads according to the distribution of unique mapped reads in mutireads mapped regions [178-179]. The solution of this problem is setting the different threshold of mutireads in my studies for different species (described in “Materials and methods” section). But these are not the best way for handling this problem. Recently, the new algorithm is designed for solving mutireads problem [177]. The algorithm is called as EM (Expectation Maximization) algorithm based on maximum likelihood approach. This mode assumes the sequencing reads are random generated from the region which has the corresponding distribution. This model originally is designed for RNA-seq which is used to monitor the expression level of mRNAs. miRNAkey, the software for miRNA analysis in NGS data, implements this EM algorithm. So, this model can be applied for optimizing the miRNA expression level (solving mutireads problem).

In profiling the expression level of miRNAs, various sequencing reads for the miRNAs (isomiRs) are always observed. Figure 42 gives the example of isomiRs in human and *Arabidopsis*. The sequence of hsa-miR-517b in miRBase is "TCGTGCATCCCTTTAGAGTGTT". The variants are shifting several nucleotides in 5' or 3' end (comparing with hsa-miR-517b). For example, "Reads 2" shifts two nucleotides in 5' end. "Reads 7" shifts one nucleotide in 5' end. The length of these variants is the same or different with hsa-miR-517b. In *Arabidopsis*, isomiRs are also found. The sequence of ath-miR775 is "TTCGATGTCTAGCAGTGCCAA". "Reads 8" and "Reads 12" shifts two and one nucleotides in 5' end, respectively. In previous studies, the highest expressed read is considered as the expression of the miRNA. So, the expression level of has-miR-517b is 377 in human. In *Arabidopsis*, the expression level of ath-miR775 is 719. This profiling method is not good enough to correctly evaluate the real expression level of miRNAs. In my study, the expression level of the miRNA is counting all isomiRs. This method is often used now. However, the tolerance range of shifted nucleotide is not defined clearly and the method for modeling isomiRs is not purposed. In addition, the sequence of highest expressed read is different with the know miRNA in two examples (Figure 32). This condition is usually observed in many small RNA NGS data. It affects the miRNA target site prediction in animals and plants. For example, the highest expressed read is shifting one nucleotide in 5' end in hsa-miR-517b. In target prediction model of animals, the seed region, position 2-8 of the miRNA, is the major factor which determines the target site. The seed region of hsa-miR-517b is changed from “CGTGCAT” to “TCGTGCA”. So, the predicted target sites are also changed. In plants, the core region of target site region is position 3-12. Therefore, the core region changing due to isomiRs certainly leads to the change of predicted sites. There are no related studies about this problem and its corresponding solution. In my study, the top three highest expressed reads are used to predict the target sites. Then, joint target sites are found from three predicted target site sets. It is not the best way for determining miRNA targets. The better method needs to be purposed in the future.



Human

miRNA	hsa-miR-517b	
miRNA	TCGTGCATCCCTTTAGAGTGTT	Read count
Reads 0	TCGTGCATCCCTTTAGAGTG	2
Reads 1	TCGTGCATCCCTTTAGAGTGT	14
Reads 2	GATCGTGCATCCCTTTAGAGTGT	1
Reads 3	ATCGTGCATCCCTTTAGAGTGTT	1
Reads 4	TCGTGCATCCCTTTAGAGTGTT	1
Reads 5	TCGTGCATCCCTTTAGAGT	1
Reads 6	ATCGTGCATCCCTTTAGAGTG	19
Reads 7	ATCGTGCATCCCTTTAGAGTGT	377
Reads 8	GTGCATCCCTTTAGAGTGT	3
Reads 9	ATCGTGCATCCCTTTAGAGT	7

Arabidopsis

miRNA	MIR775	
miRNA	TTCGATGTCTAGCAGTGCCAA	Read count
Reads 0	TTCGATGTCTAGCAGTGCCA	719
Reads 1	CGATGTCTAGCAGTGCCAA	16
Reads 2	GATGTCTAGCAGTGCCAAT	6
Reads 3	TTCGATGTCTAGCAGTGCC	1
Reads 4	TTCGATGTCTAGCAGTGCCAAT	266
Reads 5	TTTTCGATGTCTAGCAGTGCCAAT	1
Reads 6	TTCGATGTCTAGCAGTGCCAA	638
Reads 7	TTTTCGATGTCTAGCAGTGCC	8
Reads 8	TTTTCGATGTCTAGCAGTGCCAA	347
Reads 9	TTTCGATGTCTAGCAGTGCCA	5
Reads 10	TTTTCGATGTCTAGCAGTGCCA	38
Reads 11	TCGATGTCTAGCAGTGCCAA	17
Reads 12	TTTCGATGTCTAGCAGTGCCAA	33

Figure 32. The example of isomiRs in Human and *Arabidopsis*

6.2 Improving completeness of regulation role of small non-coding RNAs by adding other kind NGS data

Next-generation sequencing technology is the good way for analyzing small non-coding RNAs. To deeply investigate the regulation role of small RNAs, multiple information such as gene expression data from cDNA microarray, gene function (GO) and pathways are combined in previous studies or my study. However, it is still not enough to completely decipher the pathway of small non-coding RNAs. Several kinds of NGS data can be merged for doing more comprehensive analysis.

The first is RNA-seq which is the application of NGS. It is used to detect the expression level of transcripts. Unlike cDNA microarray, RNA-seq can detect novel alternative splicing events and more correctly monitor the expression of transcripts. The probe sets of cDNA microarray are designed based on whole known isoforms. So, RNA-seq can more sensitively detect the expression level of each isoform. In miRNAs and siRNAs target site prediction, the more correct mRNA expression profile is helpful for reducing the false positive rate of prediction due to the miRNAs and siRNAs down-regulating the expression level of mRNAs. Moreover, the mRNA expression profile of RNA-seq can also reduce the false positive rate of predicting transcriptional cis-regulatory elements in the promoter regions. For example, E2F3, a transcription factor, is predicted that induces a set of miRNAs by binding their promoter regions. If E2F3 is not detected in the mRNA expression profile, it is not the candidate for regulating these miRNAs. If the set of miRNAs are up-regulated but E2F3 is down-regulated, E2F3 is not the regulator of these miRNAs.

The second is chip (chromatin immunoprecipitation)-seq which is another application of NGS. Chip-seq is applied for high-throughput screening the DNA sequences which are bound by selected proteins. These proteins can be transcription factors (TFs) or other important proteins in biological processes. After mapping the sequencing reads of chip-seq of selected transcription factor to genomes, the TFBSs (transcription factor binding sites) are identified in the sequence abundant regions. These experimental evidences can be used to build the relationship between miRNAs and the TFs which regulate them. For example, the sequence abundant regions locate the promoter region of miRNAs. The TFs have high possibility regulating miRNAs. The complete miRNA regulatory network from transcription to translation level can be generated by combing RNA-seq and chip-seq.

The third is CLIP (crosslinking immunoprecipitation)-seq which is the novel application of NGS. CLIP-seq is the method for high-throughput screening the RNA sequences which are bound by selected proteins. Recent studies use AGO (Argonaute) protein to do CLIP-seq [180-181]. AGO protein is required for miRNAs and siRNAs targeting mRNAs. So, the miRNA and mRNA interaction sites can be detected by AGO CLIP-seq. AGO CLIP-seq can directly give the evidence of miRNA and mRNA interactions. But these interaction sites need to be further analyzed. This is because AGO CLIP-seq is global screening the miRNA and mRNA interaction sequences. Identifying what miRNAs interact with these sequences is required. For example, the interaction region is chr1: 156770-156798 [+] in human. All known miRNAs are used to find the interaction in this region or not by current target site prediction guideline.

The fourth is Degradome-Seq (degradome sequencing) which is designed for identifying the miRNA cleavage sites by using a modified 5'-rapid amplification of cDNA ends (RACE) with next-generation sequencing technology. Degradome-Seq first is used to find miRNA target sites in plants [71, 182-185]. Recently, some studies are purposed that they used Degradome-Seq to identify miRNA-derived cleavage sites [186-187]. Therefore, Degradome-Seq can give the direct evidence of miRNA and mRNA interactions in plants and animals.

The final is high-throughput sequencing of DNA methylation. There are three high-throughput sequencing methods for DNA methylation such as bisulfate-sequencing (BS-seq), methylated DNA immunoprecipitation sequencing (MeDIP-seq) and methyl-binding protein sequencing (MBD-seq) [188]. The degree of DNA methylation affects the transcription processes. In plants, the function of hc-siRNAs regulates gene expression by triggering DNA methylation. Therefore, the DNA methylation profile can be applied for identifying relationship between hc-siRNAs and their target genes.

References

1. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
2. Kent, W.J., *BLAT--the BLAST-like alignment tool.* Genome Res, 2002. **12**(4): p. 656-64.
3. Jiang, H. and W.H. Wong, *SeqMap: mapping massive amount of oligonucleotides to the genome.* Bioinformatics, 2008. **24**(20): p. 2395-6.
4. Lin, H., et al., *ZOOM! Zillions of oligos mapped.* Bioinformatics, 2008. **24**(21): p. 2431-7.
5. Li, H., J. Ruan, and R. Durbin, *Mapping short DNA sequencing reads and calling variants using mapping quality scores.* Genome Res, 2008. **18**(11): p. 1851-8.
6. Langmead, B., et al., *Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.* Genome Biol, 2009. **10**(3): p. R25.
7. Li, R., et al., *SOAP: short oligonucleotide alignment program.* Bioinformatics, 2008. **24**(5): p. 713-4.
8. Friedlander, M.R., et al., *Discovering microRNAs from deep sequencing data using miRDeep.* Nat Biotechnol, 2008. **26**(4): p. 407-15.
9. Wang, W.C., et al., *miRExpress: analyzing high-throughput sequencing data for profiling microRNA expression.* BMC Bioinformatics, 2009. **10**: p. 328.
10. Ronen, R., et al., *miRNAkey: a software for microRNA deep sequencing analysis.* Bioinformatics, 2010. **26**(20): p. 2615-6.
11. Hackenberg, M., et al., *miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments.* Nucleic Acids Res, 2009. **37**(Web Server issue): p. W68-76.
12. Pantano, L., X. Estivill, and E. Marti, *SeqBuster, a bioinformatic tool for the processing and analysis of small RNAs datasets, reveals ubiquitous miRNA modifications in human embryonic cells.* Nucleic Acids Res, 2010. **38**(5): p. e34.
13. Huang, P.J., et al., *DSAP: deep-sequencing small RNA analysis pipeline.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W385-91.
14. Zhu, E., et al., *mirTools: microRNA profiling and discovery based on high-throughput sequencing.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W392-7.
15. Kanehisa, M., *The KEGG database.* Novartis Found Symp, 2002. **247**: p. 91-101; discussion 101-3, 119-28, 244-52.
16. Berezikov, E., et al., *Diversity of microRNAs in human and chimpanzee brain.*

Nat Genet, 2006. **38**(12): p. 1375-7.

17. Morin, R.D., et al., *Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells.* Genome Res, 2008. **18**(4): p. 610-21.
18. Bar, M., et al., *MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries.* Stem Cells, 2008. **26**(10): p. 2496-505.
19. Nygaard, S., et al., *Identification and analysis of miRNAs in human breast cancer and teratoma samples using deep sequencing.* BMC Med Genomics, 2009. **2**: p. 35.
20. Jima, D.D., et al., *Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs.* Blood, 2010. **116**(23): p. e118-27.
21. Pantaleo, V., et al., *Deep sequencing analysis of viral short RNAs from an infected Pinot Noir grapevine.* Virology, 2010. **408**(1): p. 49-56.
22. Marti, E., et al., *A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing.* Nucleic Acids Res, 2010. **38**(20): p. 7219-35.
23. Ribeiro-dos-Santos, A., et al., *Ultra-deep sequencing reveals the microRNA expression pattern of the human stomach.* PLoS One, 2010. **5**(10): p. e13205.
24. Xu, M.J., et al., *Identification and characterization of microRNAs in Clonorchis sinensis of human health significance.* BMC Genomics, 2010. **11**: p. 521.
25. Shao, N.Y., et al., *Comprehensive survey of human brain microRNA by deep sequencing.* BMC Genomics, 2010. **11**: p. 409.
26. Liao, J.Y., et al., *Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly complex subcellular distribution of miRNAs and tRNA 3' trailers.* PLoS One, 2010. **5**(5): p. e10563.
27. Szczyrba, J., et al., *The microRNA profile of prostate carcinoma obtained by deep sequencing.* Mol Cancer Res, 2010. **8**(4): p. 529-38.
28. Schulte, J.H., et al., *Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma.* Nucleic Acids Res, 2010. **38**(17): p. 5919-28.
29. Creighton, C.J., et al., *Discovery of novel microRNAs in female reproductive tract using next generation sequencing.* PLoS One, 2010. **5**(3): p. e9637.
30. Xu, G., et al., *Characterization of the small RNA transcriptomes of androgen dependent and independent prostate cancer cell line by deep sequencing.* PLoS One, 2010. **5**(11): p. e15519.
31. Fehniger, T.A., et al., *Next-generation sequencing identifies the natural killer*

cell microRNA transcriptome. Genome Res, 2010. **20**(11): p. 1590-604.

32. Beck, D., et al., *Integrative analysis of next generation sequencing for small non-coding RNAs and transcriptional regulation in Myelodysplastic Syndromes.* BMC Med Genomics, 2011. **4**: p. 19.
33. Persson, H., et al., *Identification of new microRNAs in paired normal and tumor breast tissue suggests a dual role for the ERBB2/Her2 gene.* Cancer Res, 2011. **71**(1): p. 78-86.
34. Vaz, C., et al., *Analysis of microRNA transcriptome by deep sequencing of small RNA libraries of peripheral blood.* BMC Genomics, 2010. **11**: p. 288.
35. Hackl, M., et al., *Next-generation sequencing of the Chinese hamster ovary microRNA transcriptome: Identification, annotation and profiling of microRNAs as targets for cellular engineering.* J Biotechnol, 2011.
36. Ling, K.H., et al., *Deep sequencing analysis of the developing mouse brain reveals a novel microRNA.* BMC Genomics, 2011. **12**(1): p. 176.
37. Ruby, J.G., et al., *Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans.* Cell, 2006. **127**(6): p. 1193-207.
38. Stark, A., et al., *Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures.* Nature, 2007. **450**(7167): p. 219-32.
39. Burnside, J., et al., *Deep sequencing of chicken microRNAs.* BMC Genomics, 2008. **9**: p. 185.
40. Glazov, E.A., et al., *A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach.* Genome Res, 2008. **18**(6): p. 957-64.
41. Chen, X., et al., *Identification and characterization of novel amphioxus microRNAs by Solexa sequencing.* Genome Biol, 2009. **10**(7): p. R78.
42. Friedlander, M.R., et al., *High-resolution profiling and discovery of planarian small RNAs.* Proc Natl Acad Sci U S A, 2009. **106**(28): p. 11546-51.
43. Rathjen, T., et al., *High throughput sequencing of microRNAs in chicken somites.* FEBS Lett, 2009. **583**(9): p. 1422-6.
44. Jagadeeswaran, G., et al., *Deep sequencing of small RNA libraries reveals dynamic regulation of conserved and novel microRNAs and microRNA-stars during silkworm development.* BMC Genomics, 2010. **11**: p. 52.
45. Legeai, F., et al., *Bioinformatic prediction, deep sequencing of microRNAs and expression analysis during phenotypic plasticity in the pea aphid, Acyrthosiphon pisum.* BMC Genomics, 2010. **11**: p. 281.
46. Huang, Q.X., et al., *MicroRNA discovery and analysis of pinewood nematode Bursaphelenchus xylophilus by deep sequencing.* PLoS One, 2010. **5**(10): p.

e13271.

47. Chen, X., et al., *Next-generation small RNA sequencing for microRNAs profiling in the honey bee Apis mellifera.* Insect Mol Biol, 2010. **19**(6): p. 799-805.

48. Li, S.C., et al., *Discovery and characterization of medaka miRNA genes by next generation sequencing platform.* BMC Genomics, 2010. **11 Suppl 4**: p. S8.

49. Wei, Z., et al., *Novel and conserved micrornas in dalian purple urchin (strongylocentrotus nudus) identified by next generation sequencing.* Int J Biol Sci, 2011. **7**(2): p. 180-92.

50. Wang, X.J., T. Gaasterland, and N.H. Chua, *Genome-wide prediction and identification of cis-natural antisense transcripts in Arabidopsis thaliana.* Genome Biol, 2005. **6**(4): p. R30.

51. Rajagopalan, R., et al., *A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana.* Genes Dev, 2006. **20**(24): p. 3407-25.

52. Fahlgren, N., et al., *High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes.* PLoS One, 2007. **2**(2): p. e219.

53. Kasschau, K.D., et al., *Genome-wide profiling and analysis of Arabidopsis siRNAs.* PLoS Biol, 2007. **5**(3): p. e57.

54. Mi, S., et al., *Sorting of small RNAs into Arabidopsis argonaute complexes is directed by the 5' terminal nucleotide.* Cell, 2008. **133**(1): p. 116-27.

55. Hsieh, L.C., et al., *Uncovering small RNA-mediated responses to phosphate deficiency in Arabidopsis by deep sequencing.* Plant Physiol, 2009. **151**(4): p. 2120-32.

56. Qi, X., F.S. Bao, and Z. Xie, *Small RNA deep sequencing reveals role for Arabidopsis thaliana RNA-dependent RNA polymerases in viral siRNA biogenesis.* PLoS One, 2009. **4**(3): p. e4971.

57. Zhang, W., et al., *Bacteria-responsive microRNAs regulate plant innate immunity by modulating plant hormone networks.* Plant Mol Biol, 2011. **75**(1-2): p. 93-105.

58. Sunkar, R., et al., *Identification of novel and candidate miRNAs in rice by high throughput sequencing.* BMC Plant Biol, 2008. **8**: p. 25.

59. Lu, C., et al., *Genome-wide analysis for discovery of rice microRNAs reveals natural antisense microRNAs (nat-miRNAs).* Proc Natl Acad Sci U S A, 2008. **105**(12): p. 4951-6.

60. Zhou, X., et al., *Genome-wide identification and analysis of small RNAs originated from natural antisense transcripts in Oryza sativa.* Genome Res, 2009. **19**(1): p. 70-8.

61. Nakano, M., et al., *Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA.* Nucleic Acids Res, 2006. **34**(Database issue): p. D731-5.

62. Yao, Y., et al., *Cloning and characterization of microRNAs from wheat (Triticum aestivum L.).* Genome Biol, 2007. **8**(6): p. R96.

63. Szittya, G., et al., *High-throughput sequencing of Medicago truncatula short RNAs identifies eight new miRNA families.* BMC Genomics, 2008. **9**: p. 593.

64. Moxon, S., et al., *Deep sequencing of tomato short RNAs identifies microRNAs targeting genes involved in fruit ripening.* Genome Res, 2008. **18**(10): p. 1602-9.

65. Mica, E., et al., *High throughput approaches reveal splicing of primary microRNA transcripts and tissue specific expression of mature microRNAs in Vitis vinifera.* BMC Genomics, 2009. **10**: p. 558.

66. Wei, B., et al., *Novel microRNAs uncovered by deep sequencing of small RNA transcriptomes in bread wheat (Triticum aestivum L.) and Brachypodium distachyon (L.) Beauv.* Funct Integr Genomics, 2009. **9**(4): p. 499-511.

67. Zhang, L., et al., *A genome-wide characterization of microRNA genes in maize.* PLoS Genet, 2009. **5**(11): p. e1000716.

68. Zhang, J., et al., *Deep sequencing of Brachypodium small RNAs at the global genome level identifies microRNAs involved in cold stress response.* BMC Genomics, 2009. **10**: p. 449.

69. Zhao, C.Z., et al., *Deep sequencing identifies novel and conserved microRNAs in peanuts (Arachis hypogaea L.).* BMC Plant Biol, 2010. **10**: p. 3.

70. Song, C., et al., *Deep sequencing discovery of novel and conserved microRNAs in trifoliate orange (Citrus trifoliata).* BMC Genomics, 2010. **11**: p. 431.

71. Pantaleo, V., et al., *Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis.* Plant J, 2010. **62**(6): p. 960-76.

72. Martinez, G., et al., *High-throughput sequencing of Hop stunt viroid-derived small RNAs from cucumber leaves and phloem.* Mol Plant Pathol, 2010. **11**(3): p. 347-59.

73. Song, Q.X., et al., *Identification of miRNAs and their target genes in developing soybean seeds by deep sequencing.* BMC Plant Biol, 2011. **11**: p. 5.

74. Lu, Y.C., et al., *Deep sequencing identifies new and regulated microRNAs in Schmidtea mediterranea.* RNA, 2009. **15**(8): p. 1483-91.

75. Huang, J., et al., *Genome-wide identification of Schistosoma japonicum microRNAs using a deep-sequencing approach.* PLoS One, 2009. **4**(12): p. e8206.

76. Chen, X.S., et al., *High throughput genome-wide survey of small RNAs from the parasitic protists Giardia intestinalis and Trichomonas vaginalis.* Genome Biol Evol, 2009. **1**: p. 165-75.
77. Irnov, I., et al., *Identification of regulatory RNAs in Bacillus subtilis.* Nucleic Acids Res, 2010. **38**(19): p. 6637-51.
78. Donaire, L., et al., *Deep-sequencing of plant viral small RNAs reveals effective and widespread targeting of viral genomes.* Virology, 2009. **392**(2): p. 203-14.
79. Zhu, J.Y., et al., *Identification and analysis of expression of novel microRNAs of murine gammaherpesvirus 68.* J Virol, 2010. **84**(19): p. 10266-75.
80. Varshney, R.K., et al., *Next-generation sequencing technologies and their implications for crop genetics and breeding.* Trends Biotechnol, 2009. **27**(9): p. 522-30.
81. Turner, D.J., et al., *Next-generation sequencing of vertebrate experimental organisms.* Mamm Genome, 2009. **20**(6): p. 327-38.
82. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.
83. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Hum Mol Genet, 2010. **19**(R2): p. R227-40.
84. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation.* Nat Rev Genet, 2004. **5**(7): p. 522-31.
85. Esquela-Kerscher, A. and F.J. Slack, *Oncomirs - microRNAs with a role in cancer.* Nat Rev Cancer, 2006. **6**(4): p. 259-69.
86. Denli, A.M., et al., *Processing of primary microRNAs by the Microprocessor complex.* Nature, 2004. **432**(7014): p. 231-5.
87. Ketting, R.F., et al., *Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans.* Genes Dev, 2001. **15**(20): p. 2654-9.
88. Meister, G., et al., *Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs.* Mol Cell, 2004. **15**(2): p. 185-97.
89. Ohrt, T., et al., *Fluorescence correlation spectroscopy and fluorescence cross-correlation spectroscopy reveal the cytoplasmic origination of loaded nuclear RISC in vivo in human cells.* Nucleic Acids Res, 2008. **36**(20): p. 6439-49.
90. Diederichs, S. and D.A. Haber, *Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression.* Cell, 2007. **131**(6): p. 1097-108.
91. Grimson, A., et al., *MicroRNA targeting specificity in mammals: determinants beyond seed pairing.* Mol Cell, 2007. **27**(1): p. 91-105.

92. Selbach, M., et al., *Widespread changes in protein synthesis induced by microRNAs.* Nature, 2008. **455**(7209): p. 58-63.

93. Wang, X., et al., *Aberrant expression of oncogenic and tumor-suppressive microRNAs in cervical cancer is required for cancer cell growth.* PLoS One, 2008. **3**(7): p. e2557.

94. Yu, B., et al., *Methylation as a crucial step in plant microRNA biogenesis.* Science, 2005. **307**(5711): p. 932-5.

95. Vaucheret, H., *Post-transcriptional small RNA pathways in plants: mechanisms and regulations.* Genes Dev, 2006. **20**(7): p. 759-71.

96. Papp, I., et al., *Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors.* Plant Physiol, 2003. **132**(3): p. 1382-90.

97. Reinhart, B.J., et al., *MicroRNAs in plants.* Genes Dev, 2002. **16**(13): p. 1616-26.

98. Park, W., et al., *CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana.* Curr Biol, 2002. **12**(17): p. 1484-95.

99. Park, M.Y., et al., *Nuclear processing and export of microRNAs in Arabidopsis.* Proc Natl Acad Sci U S A, 2005. **102**(10): p. 3691-6.

100. Bartel, D.P., *MicroRNAs: genomics, biogenesis, mechanism, and function.* Cell, 2004. **116**(2): p. 281-97.

101. Kim, V.N., *Small RNAs: classification, biogenesis, and function.* Mol Cells, 2005. **19**(1): p. 1-15.

102. Du, T. and P.D. Zamore, *microPrimer: the biogenesis and function of microRNA.* Development, 2005. **132**(21): p. 4645-52.

103. Ambros, V., *The functions of animal microRNAs.* Nature, 2004. **431**(7006): p. 350-5.

104. Palatnik, J.F., et al., *Control of leaf morphogenesis by microRNAs.* Nature, 2003. **425**(6955): p. 257-63.

105. Mallory, A.C., et al., *MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region.* EMBO J, 2004. **23**(16): p. 3356-64.

106. Jones-Rhoades, M.W. and D.P. Bartel, *Computational identification of plant microRNAs and their targets, including a stress-induced miRNA.* Mol Cell, 2004. **14**(6): p. 787-99.

107. Rhoades, M.W., et al., *Prediction of plant microRNA targets.* Cell, 2002. **110**(4): p. 513-20.

108. Sunkar, R., et al., *Small RNAs as big players in plant abiotic stress responses and nutrient deprivation.* Trends Plant Sci, 2007. **12**(7): p. 301-9.

109. Sunkar, R., A. Kapoor, and J.K. Zhu, *Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in Arabidopsis is mediated by downregulation of miR398 and important for oxidative stress tolerance.* Plant Cell, 2006. **18**(8): p. 2051-65.

110. Navarro, L., et al., *A plant miRNA contributes to antibacterial resistance by repressing auxin signaling.* Science, 2006. **312**(5772): p. 436-9.

111. Chiou, T.J., et al., *Regulation of phosphate homeostasis by MicroRNA in Arabidopsis.* Plant Cell, 2006. **18**(2): p. 412-21.

112. Fujii, H., et al., *A miRNA involved in phosphate-starvation response in Arabidopsis.* Curr Biol, 2005. **15**(22): p. 2038-43.

113. Yoshikawa, M., et al., *A pathway for the biogenesis of trans-acting siRNAs in Arabidopsis.* Genes Dev, 2005. **19**(18): p. 2164-75.

114. Allen, E., et al., *microRNA-directed phasing during trans-acting siRNA biogenesis in plants.* Cell, 2005. **121**(2): p. 207-21.

115. Vazquez, F., et al., *Endogenous trans-acting siRNAs regulate the accumulation of Arabidopsis mRNAs.* Mol Cell, 2004. **16**(1): p. 69-79.

116. Bonnet, E., Y. Van de Peer, and P. Rouze, *The small RNA world of plants.* New Phytol, 2006. **171**(3): p. 451-68.

117. Jin, H., et al., *Small RNAs and the regulation of cis-natural antisense transcripts in Arabidopsis.* BMC Mol Biol, 2008. **9**: p. 6.

118. Henz, S.R., et al., *Distinct expression patterns of natural antisense transcripts in Arabidopsis.* Plant Physiol, 2007. **144**(3): p. 1247-55.

119. Zhang, Y., et al., *Genome-wide in silico identification and analysis of cis natural antisense transcripts (cis-NATs) in ten species.* Nucleic Acids Res, 2006. **34**(12): p. 3465-75.

120. Wang, H., N.H. Chua, and X.J. Wang, *Prediction of trans-antisense transcripts in Arabidopsis thaliana.* Genome Biol, 2006. **7**(10): p. R92.

121. Steigele, S. and K. Nieselt, *Open reading frames provide a rich pool of potential natural antisense transcripts in fungal genomes.* Nucleic Acids Res, 2005. **33**(16): p. 5034-44.

122. Rosok, O. and M. Sioud, *Systematic identification of sense-antisense transcripts in mammalian cells.* Nat Biotechnol, 2004. **22**(1): p. 104-8.

123. Yelin, R., et al., *Widespread occurrence of antisense transcription in the human genome.* Nat Biotechnol, 2003. **21**(4): p. 379-86.

124. Osato, N., et al., *Antisense transcripts with rice full-length cDNAs.* Genome Biol, 2003. **5**(1): p. R5.

125. Shendure, J. and G.M. Church, *Computational discovery of sense-antisense transcription in the human and mouse genomes.* Genome Biol, 2002. **3**(9): p.

RESEARCH0044.

126. Borsani, O., et al., *Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis.* Cell, 2005. **123**(7): p. 1279-91.

127. Iida, K., et al., *Genome-wide analysis of alternative pre-mRNA splicing in Arabidopsis thaliana based on full-length cDNA sequences.* Nucleic Acids Res, 2004. **32**(17): p. 5096-103.

128. Sureau, A., et al., *Characterization of multiple alternative RNAs resulting from antisense transcription of the PR264/SC35 splicing factor gene.* Nucleic Acids Res, 1997. **25**(22): p. 4513-22.

129. Munroe, S.H. and M.A. Lazar, *Inhibition of c-erbA mRNA splicing by a naturally occurring antisense RNA.* J Biol Chem, 1991. **266**(33): p. 22083-6.

130. Sunkar, R., T. Girke, and J.K. Zhu, *Identification and characterization of endogenous small interfering RNAs from rice.* Nucleic Acids Res, 2005. **33**(14): p. 4443-54.

131. Aravin, A. and T. Tuschl, *Identification and characterization of small RNAs involved in RNA silencing.* FEBS Lett, 2005. **579**(26): p. 5830-40.

132. Xie, Z., et al., *Genetic and functional diversification of small RNA pathways in plants.* PLoS Biol, 2004. **2**(5): p. E104.

133. Chen, X., *Small RNAs and their roles in plant development.* Annu Rev Cell Dev Biol, 2009. **25**: p. 21-44.

134. Pontes, O., et al., *The Arabidopsis chromatin-modifying nuclear siRNA pathway involves a nucleolar RNA processing center.* Cell, 2006. **126**(1): p. 79-92.

135. Herr, A.J., et al., *RNA polymerase IV directs silencing of endogenous DNA.* Science, 2005. **308**(5718): p. 118-20.

136. Chan, S.W., et al., *RNA silencing genes control de novo DNA methylation.* Science, 2004. **303**(5662): p. 1336.

137. Zilberman, D., X. Cao, and S.E. Jacobsen, *ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation.* Science, 2003. **299**(5607): p. 716-9.

138. Zimmerman, A.L. and S. Wu, *MicroRNAs, cancer and cancer stem cells.* Cancer Lett, 2011. **300**(1): p. 10-9.

139. Rayner, K.J., et al., *MiR-33 contributes to the regulation of cholesterol homeostasis.* Science, 2010. **328**(5985): p. 1570-3.

140. Najafi-Shoushtari, S.H., et al., *MicroRNA-33 and the SREBP host genes cooperate to control cholesterol homeostasis.* Science, 2010. **328**(5985): p. 1566-9.

141. Horie, T., et al., *MicroRNA-33 encoded by an intron of sterol regulatory element-binding protein 2 (Srebp2) regulates HDL in vivo.* Proc Natl Acad Sci U S A, 2010. **107**(40): p. 17321-6.

142. Meng, Y., et al., *MicroRNA-mediated signaling involved in plant root development.* Biochem Biophys Res Commun, 2010. **393**(3): p. 345-9.

143. Griffiths-Jones, S., et al., *miRBase: tools for microRNA genomics.* Nucleic Acids Res, 2008. **36**(Database issue): p. D154-8.

144. Lewis, B.P., et al., *Prediction of mammalian microRNA targets.* Cell, 2003. **115**(7): p. 787-98.

145. John, B., et al., *Human MicroRNA targets.* PLoS Biol, 2004. **2**(11): p. e363.

146. Krek, A., et al., *Combinatorial microRNA target predictions.* Nat Genet, 2005. **37**(5): p. 495-500.

147. Kertesz, M., et al., *The role of site accessibility in microRNA target recognition.* Nat Genet, 2007. **39**(10): p. 1278-84.

148. Kruger, J. and M. Rehmsmeier, *RNAhybrid: microRNA target prediction easy, fast and flexible.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W451-4.

149. Yang, J.H., et al., *deepBase: a database for deeply annotating and mining deep sequencing data.* Nucleic Acids Res, 2010. **38**(Database issue): p. D123-30.

150. Gurtowski, J., et al., *Geoseq: a tool for dissecting deep-sequencing datasets.* BMC Bioinformatics, 2010. **11**: p. 506.

151. Sales, G., et al., *MAGIA, a web-based tool for miRNA and Genes Integrated Analysis.* Nucleic Acids Res, 2010. **38**(Web Server issue): p. W352-9.

152. Cho, S., et al., *miRGator v2.0: an integrated system for functional investigation of microRNAs.* Nucleic Acids Res, 2011. **39**(Database issue): p. D158-62.

153. Sethupathy, P., B. Corda, and A.G. Hatzigeorgiou, *TarBase: A comprehensive database of experimentally supported animal microRNA targets.* RNA, 2006. **12**(2): p. 192-7.

154. Xiao, F., et al., *miRecords: an integrated resource for microRNA-target interactions.* Nucleic Acids Res, 2009. **37**(Database issue): p. D105-10.

155. Jiang, Q., et al., *miR2Disease: a manually curated database for microRNA deregulation in human disease.* Nucleic Acids Res, 2009. **37**(Database issue): p. D98-104.

156. Hsu, S.D., et al., *miRTarBase: a database curates experimentally validated microRNA-target interactions.* Nucleic Acids Res, 2011. **39**(Database issue): p. D163-9.

157. Yang, J.H., et al., *starBase: a database for exploring microRNA-mRNA*

interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic Acids Res, 2011. **39**(Database issue): p. D202-9.

158. Shiraki, T., et al., *Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage.* Proc Natl Acad Sci U S A, 2003. **100**(26): p. 15776-81.

159. Barski, A., et al., *High-resolution profiling of histone methylations in the human genome.* Cell, 2007. **129**(4): p. 823-37.

160. Wakaguri, H., et al., *DBTSS: database of transcription start sites, progress report 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D97-101.

161. Wingender, E., H. Karas, and R. Knuppel, *TRANSFAC database as a bridge between sequence data libraries and biological function.* Pac Symp Biocomput, 1997: p. 477-85.

162. Portales-Casamar, E., et al., *JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2010. **38**(Database issue): p. D105-10.

163. Kel, A.E., et al., *MATCH: A tool for searching transcription factor binding sites in DNA sequences.* Nucleic Acids Res, 2003. **31**(13): p. 3576-9.

164. Gardner, P.P., et al., *Rfam: updates to the RNA families database.* Nucleic Acids Res, 2009. **37**(Database issue): p. D136-40.

165. Chang, T.H., J.T. Horng, and H.D. Huang, *RNALogo: a new approach to display structural RNA alignment.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. W91-6.

166. Warburton, P.E., et al., *Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes.* Genome Res, 2004. **14**(10A): p. 1861-9.

167. Benson, G., *Tandem repeats finder: a program to analyze DNA sequences.* Nucleic Acids Res, 1999. **27**(2): p. 573-80.

168. Lim, L.P., et al., *The microRNAs of Caenorhabditis elegans.* Genes Dev, 2003. **17**(8): p. 991-1008.

169. Lim, L.P., et al., *Vertebrate microRNA genes.* Science, 2003. **299**(5612): p. 1540.

170. Lai, E.C., et al., *Computational identification of Drosophila microRNA genes.* Genome Biol, 2003. **4**(7): p. R42.

171. Baev, V., E. Daskalova, and I. Minkov, *Computational identification of novel microRNA homologs in the chimpanzee genome.* Comput Biol Chem, 2009. **33**(1): p. 62-70.

172. Kuhn, R.M., et al., *The UCSC Genome Browser Database: update 2009.* Nucleic

Acids Res, 2009. **37**(Database issue): p. D755-61.

173. Zhang, X., et al., *Arabidopsis Argonaute 2 Regulates Innate Immunity via miRNA393(*)-Mediated Silencing of a Golgi-Localized SNARE Gene, MEMB12.* Mol Cell, 2011. **42**(3): p. 356-66.

174. Tsai, M.S., Y.H. Hsu, and N.S. Lin, *Bamboo mosaic potexvirus satellite RNA (satBaMV RNA)-encoded P20 protein preferentially binds to satBaMV RNA.* J Virol, 1999. **73**(4): p. 3032-9.

175. Lin, K.Y., et al., *Global analyses of small interfering RNAs derived from Bamboo mosaic virus and its associated satellite RNAs in different plants.* PLoS One, 2010. **5**(8): p. e11928.

176. Hsu, Y.H., et al., *Crucial role of the 5' conserved structure of bamboo mosaic virus satellite RNA in downregulation of helper viral RNA replication.* J Virol, 2006. **80**(5): p. 2566-74.

177. Pasaniuc, B., N. Zaitlen, and E. Halperin, *Accurate estimation of expression levels of homologous genes in RNA-seq experiments.* J Comput Biol, 2011. **18**(3): p. 459-68.

178. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.

179. Hashimoto, T., et al., *Probabilistic resolution of multi-mapping reads in massively parallel sequencing data using MuMRescueLite.* Bioinformatics, 2009. **25**(19): p. 2613-4.

180. Zisoulis, D.G., et al., *Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans.* Nat Struct Mol Biol, 2010. **17**(2): p. 173-9.

181. Chi, S.W., et al., *Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps.* Nature, 2009. **460**(7254): p. 479-86.

182. Wu, L., et al., *Rice MicroRNA effector complexes and targets.* Plant Cell, 2009. **21**(11): p. 3421-35.

183. Henderson, I.R. and S.E. Jacobsen, *Sequencing sliced ends reveals microRNA targets.* Nat Biotechnol, 2008. **26**(8): p. 881-2.

184. German, M.A., et al., *Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends.* Nat Biotechnol, 2008. **26**(8): p. 941-6.

185. Addo-Quaye, C., et al., *Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome.* Curr Biol, 2008. **18**(10): p. 758-62.

186. Shin, C., et al., *Expanding the microRNA targeting code: functional sites with centered pairing.* Mol Cell, 2010. **38**(6): p. 789-802.

187. Karginov, F.V., et al., *Diverse endonucleolytic cleavage sites in the mammalian transcriptome depend upon microRNAs, Drosha, and additional nucleases.*

Mol Cell, 2010. **38**(6): p. 781-8.

188. Li, N., et al., *Whole genome DNA methylation analysis based on high throughput sequencing technology.* Methods, 2010. **52**(3): p. 203-12.

Appendix

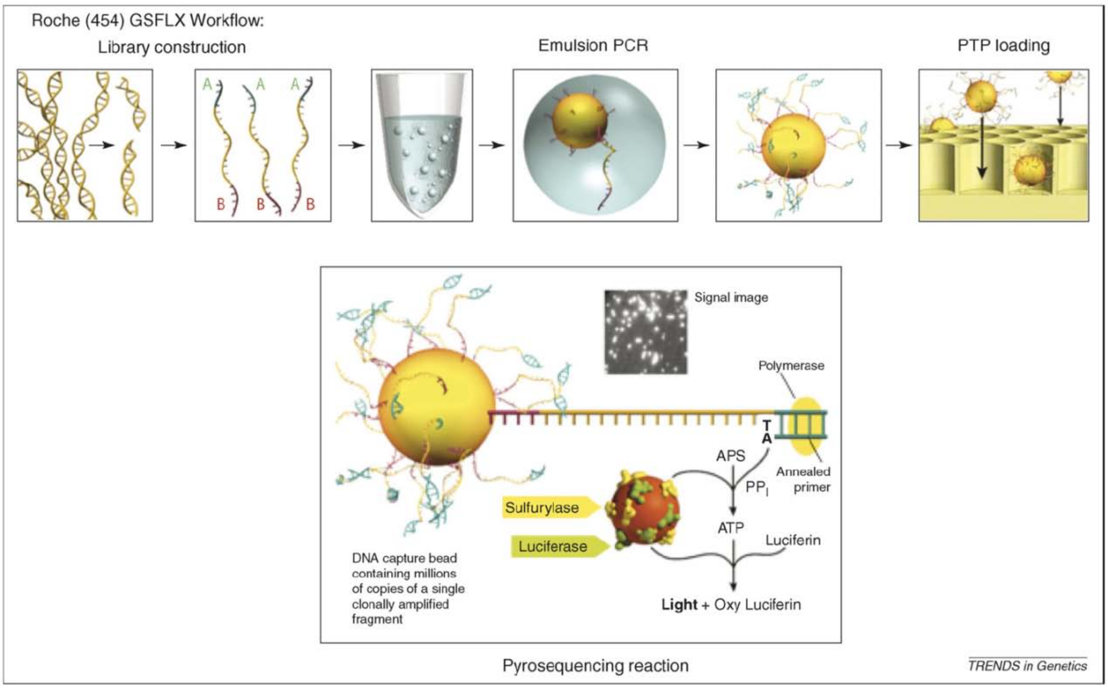
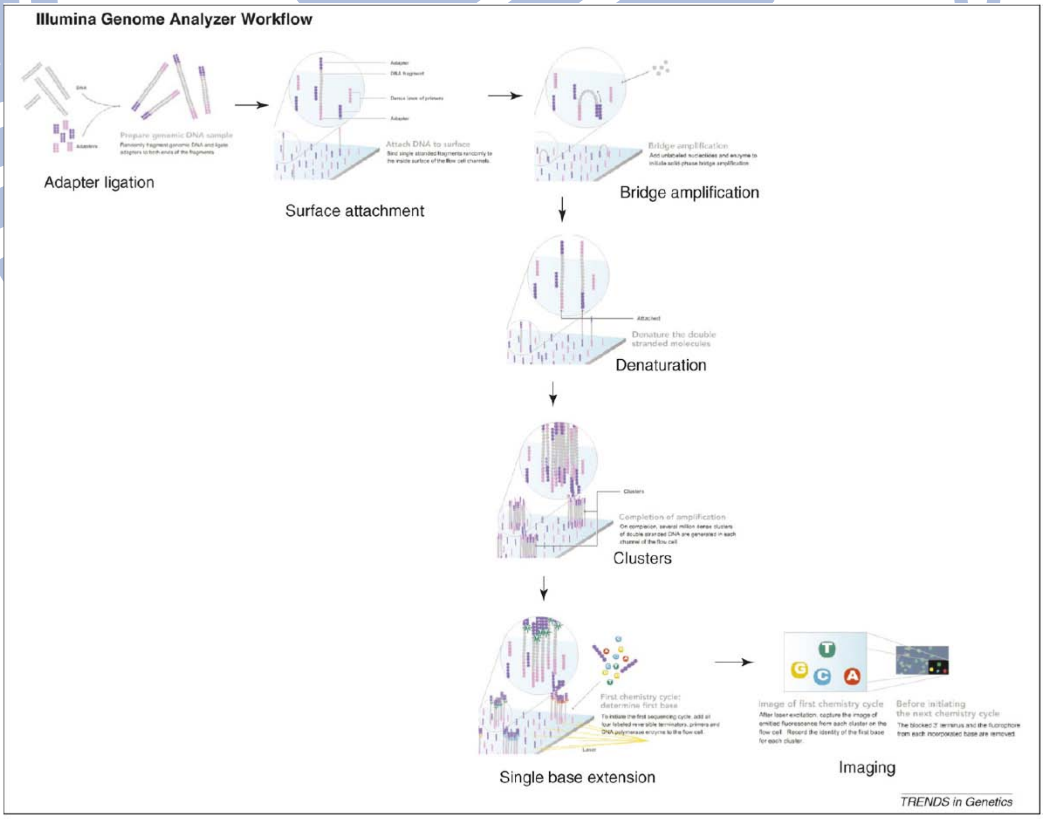


Figure S1. The workflow of Roche 454 sequencer [1]



Illumina Genome Analyzer Workflow
DNA
Adapters
Prepare genomic DNA sample
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments
Adapter ligation
Adapter
DNA fragment
Dense lawn of primers
Adapter
Attach DNA to surface
Bind single stranded fragments randomly to the inside surface of the flow cell channels
Surface attachment
Bridge amplification
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification
Bridge amplification
Attached
Denature the double stranded molecules
Denaturation
Clusters
Completion of amplification
On completion, several million dense clusters of double stranded DNA are generated in each channel of the flow cell
Clusters
First chemistry cycle: determine first base
To initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell
Laser
Single base extension
T
G
C
A
Image of first chemistry cycle
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
Before initiating the next chemistry cycle
The blocked 3' terminus and the fluorophore from each incorporated base are removed
Imaging
TRENDS in Genetics

Figure S2. The workflow of Illumia Genome Analyzer [1]

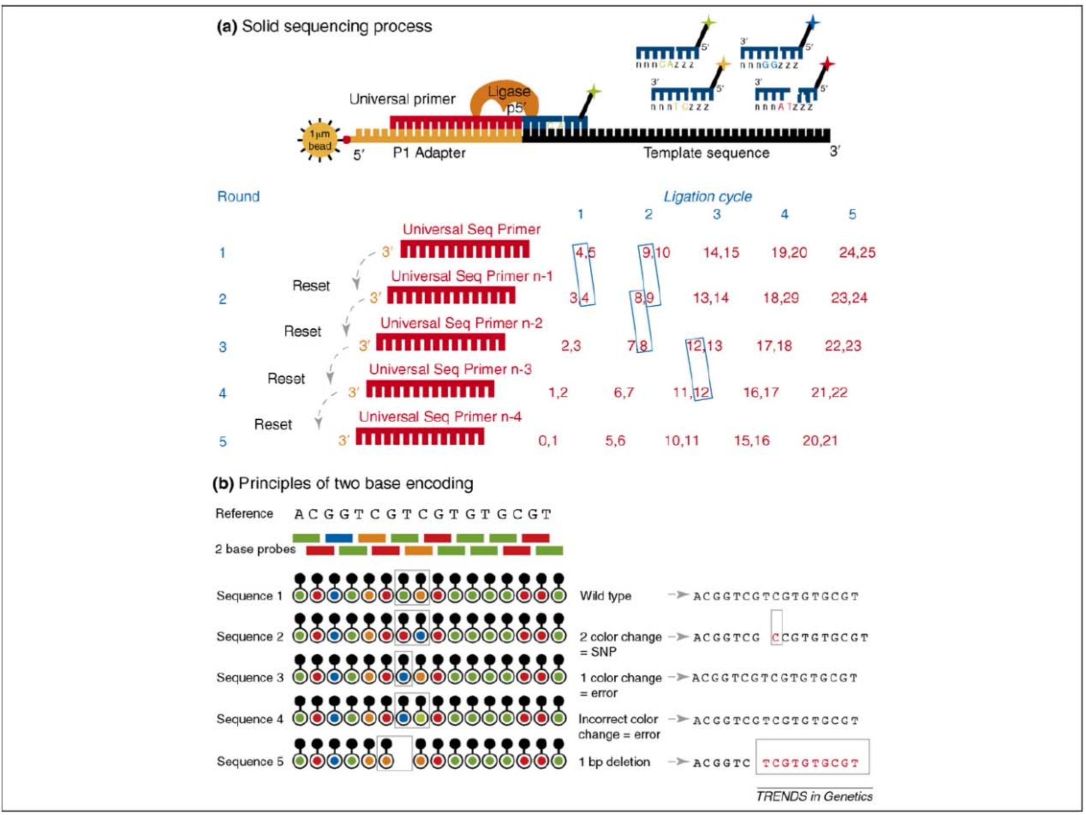


Figure S3. The workflow of ABI SOLiD sequencer [1]

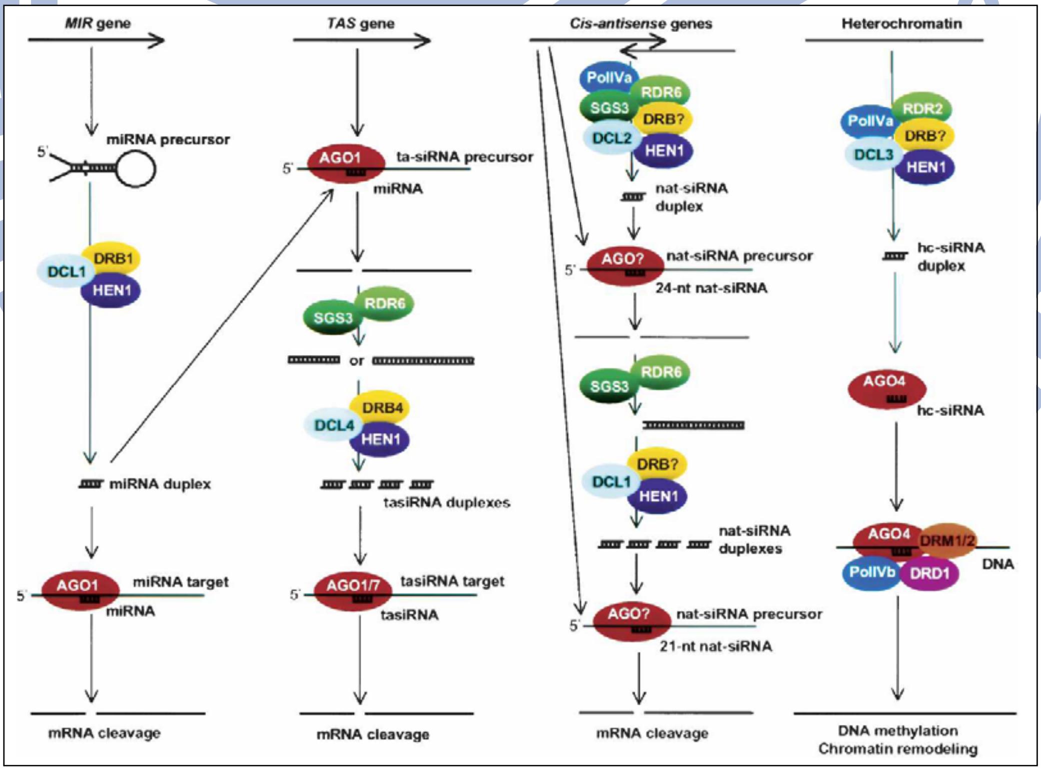


Figure S4. The biogenesis and relationship of small RNAs in plants [2]

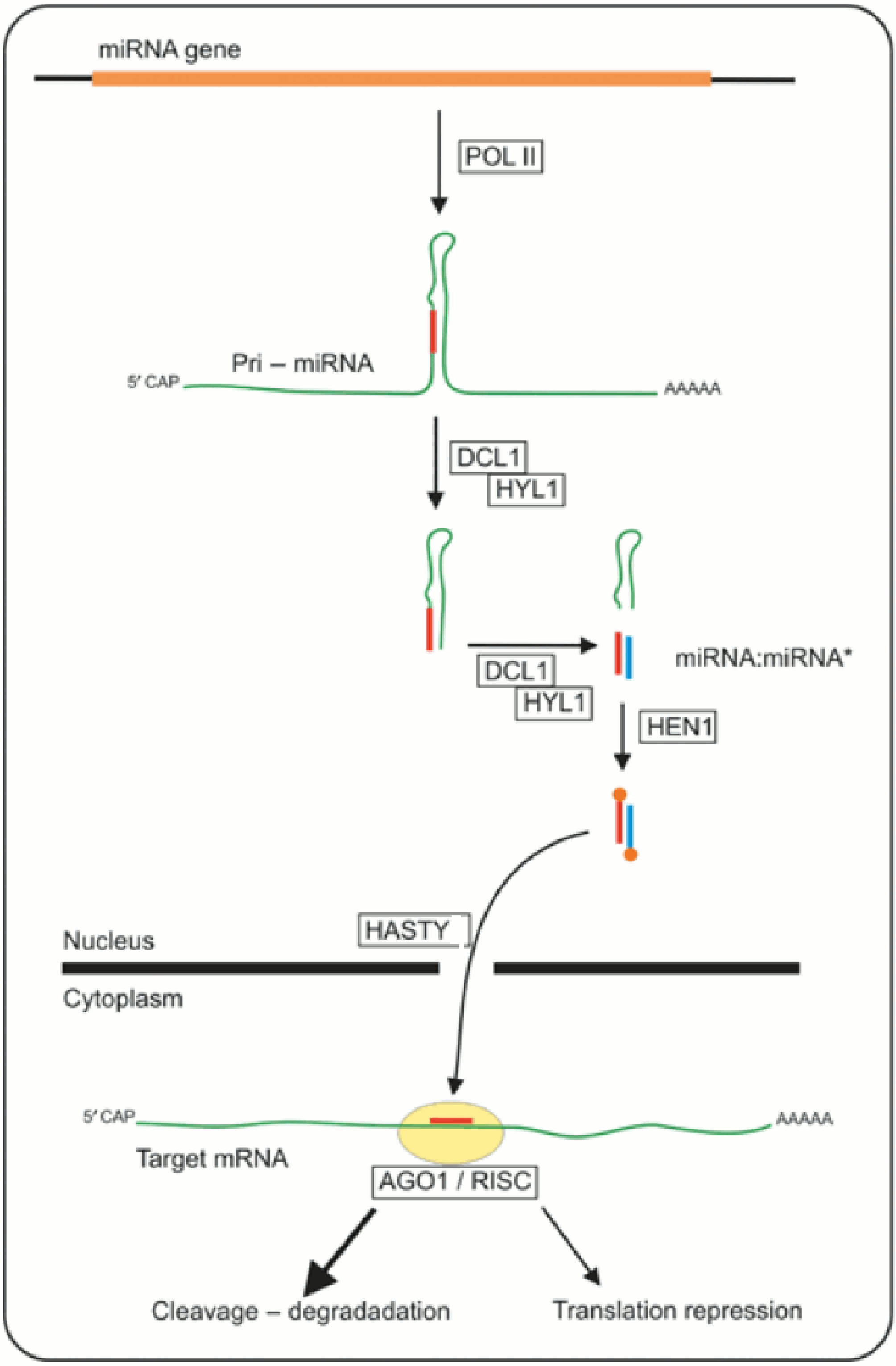


Figure S5. The biogenesis of plant miRNAs [3]

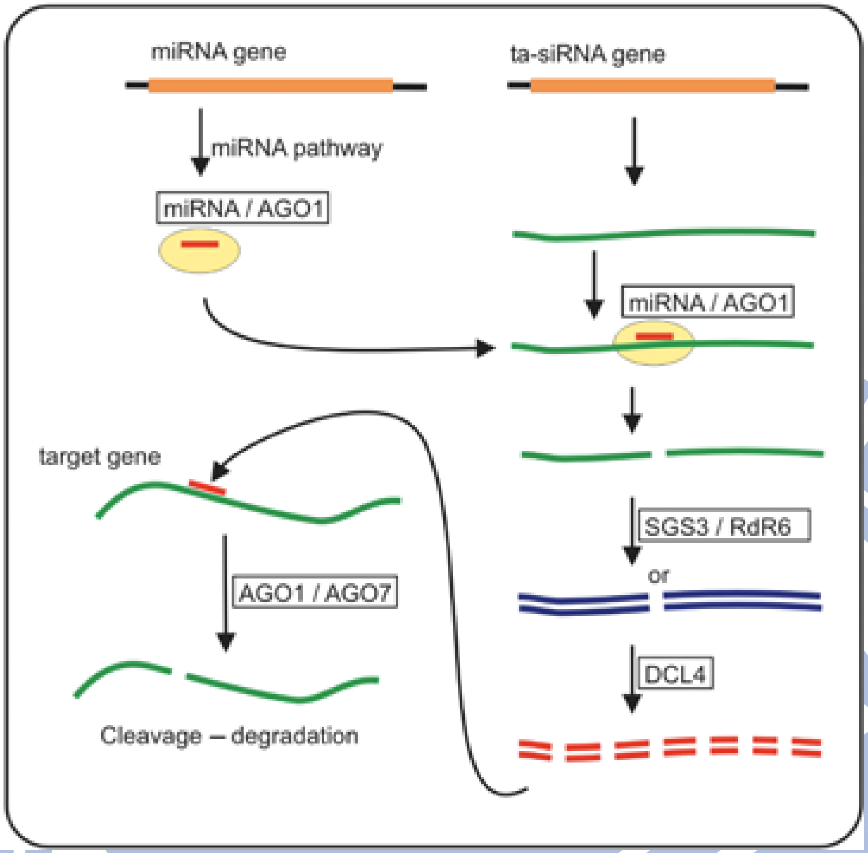


Figure S6. The biogenesis of tasiRNA [3]

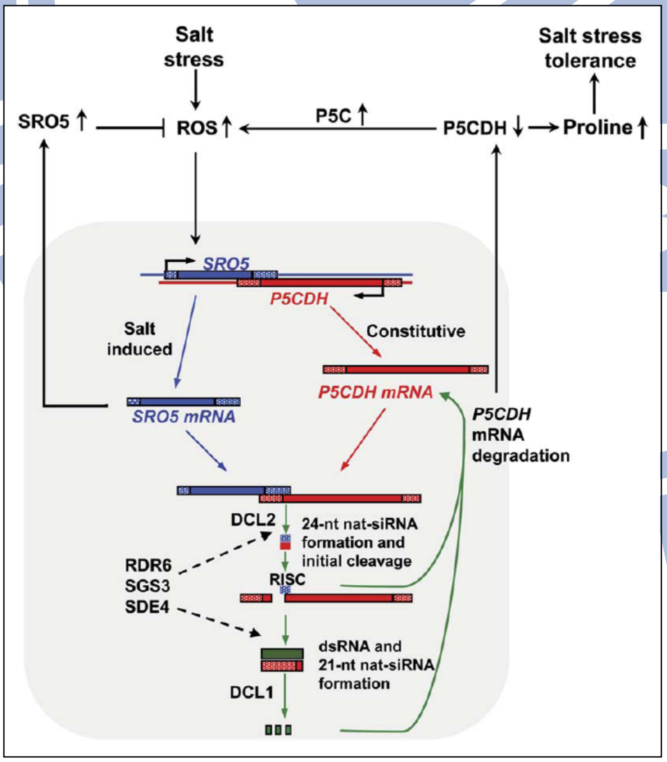


Figure S7. The regulatory network of SRO5 and P5CDH [4]

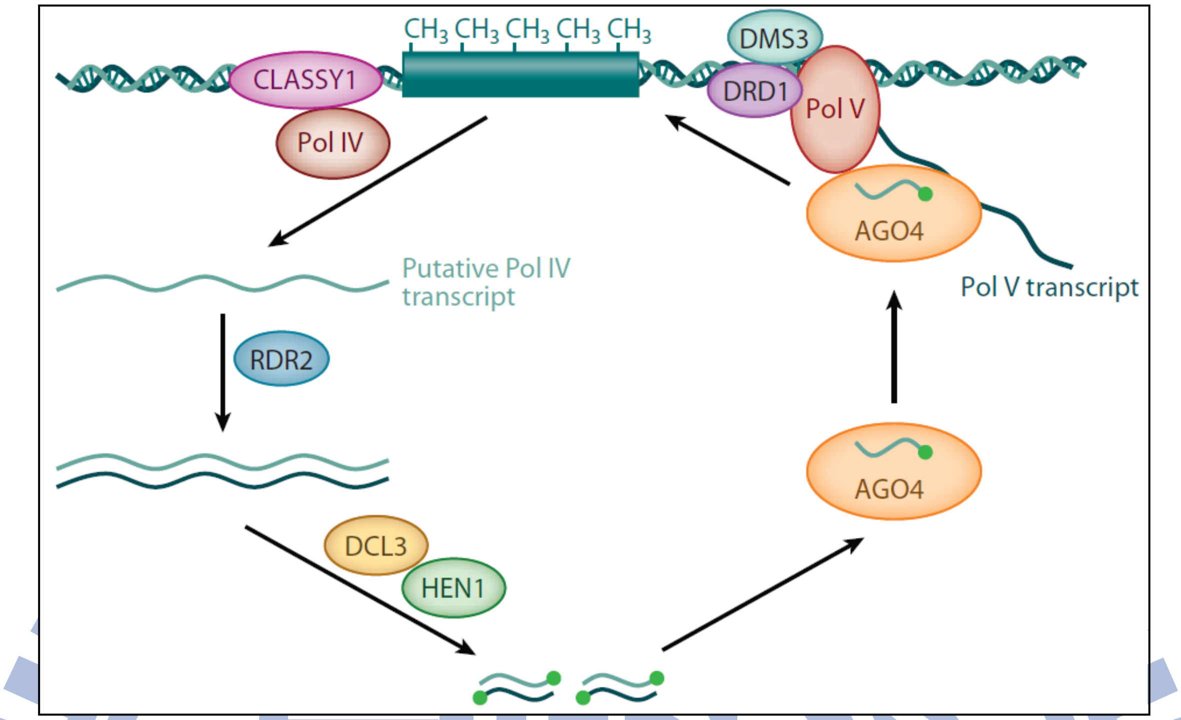


Figure S8. The biogenesis and function of heterochromatic siRNAs (hc-siRNAs) [5]

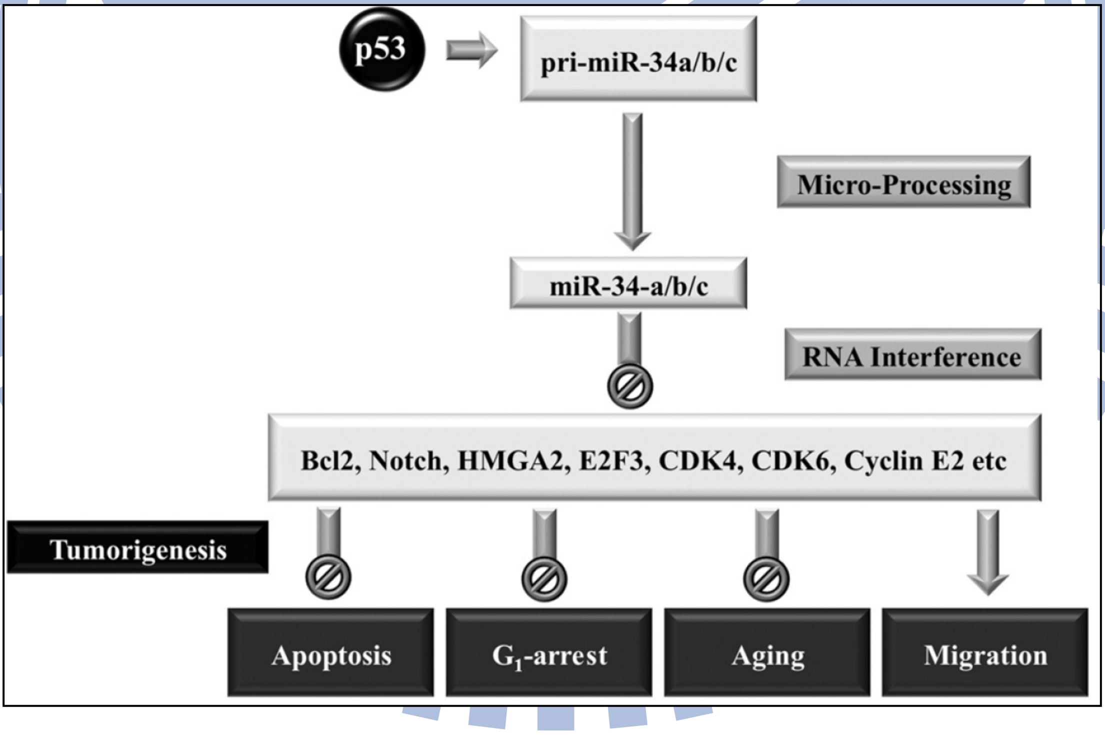


Figure S9. Regulation role of miR-34 family in cancer pathway [6]

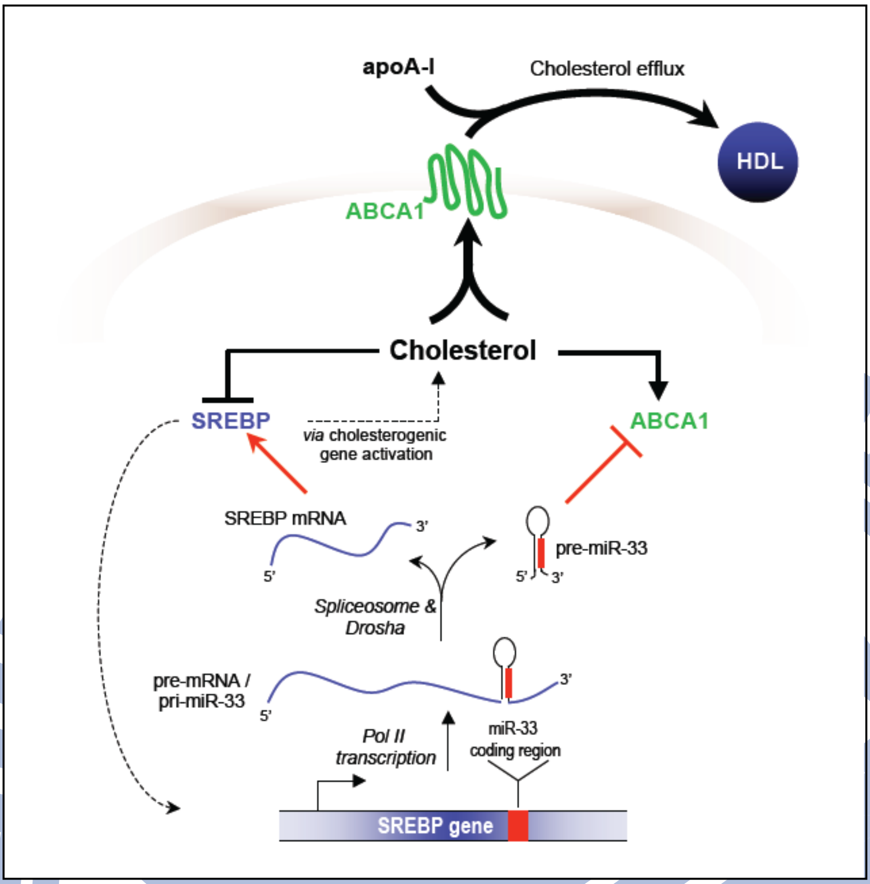
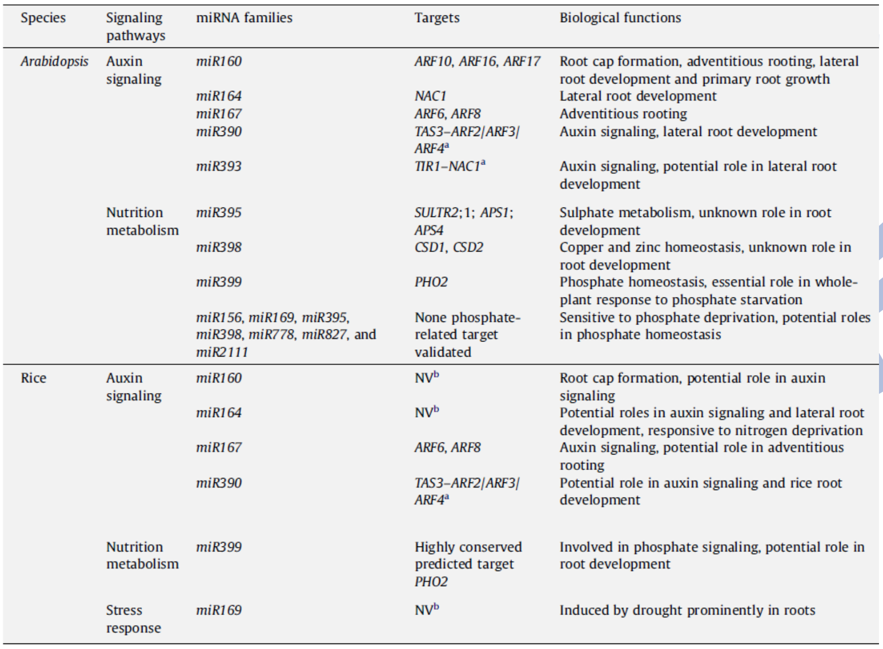


Figure S10. Regulation role of miR-33 in c cholesterol metabolic pathway [7]

Table S1. The lists of miRNAs involve in root development [8]



Species	Signaling pathways	miRNA families	Targets	Biological functions
Arabidopsis	Auxin signaling	*miR160*	*ARF10, ARF16, ARF17*	Root cap formation, adventitious rooting, lateral root development and primary root growth
		miR164	*NAC1*	Lateral root development
		miR167	*ARF6, ARF8*	Adventitious rooting
		miR390	*TAS3–ARF2/ARF3/ ARF4*[a]	Auxin signaling, lateral root development
		miR393	*TIR1–NAC1*[a]	Auxin signaling, potential role in lateral root development
	Nutrition metabolism	*miR395*	*SULTR2;1; APS1; APS4*	Sulphate metabolism, unknown role in root development
		miR398	*CSD1, CSD2*	Copper and zinc homeostasis, unknown role in root development
		miR399	*PHO2*	Phosphate homeostasis, essential role in whole-plant response to phosphate starvation
		miR156, miR169, miR395, miR398, miR778, miR827, and *miR2111*	None phosphate-related target validated	Sensitive to phosphate deprivation, potential roles in phosphate homeostasis
Rice	Auxin signaling	*miR160*	NV[b]	Root cap formation, potential role in auxin signaling
		miR164	NV[b]	Potential roles in auxin signaling and lateral root development, responsive to nitrogen deprivation
		miR167	*ARF6, ARF8*	Auxin signaling, potential role in adventitious rooting
		miR390	*TAS3–ARF2/ARF3/ ARF4*[a]	Potential role in auxin signaling and rice root development
	Nutrition metabolism	*miR399*	Highly conserved predicted target *PHO2*	Involved in phosphate signaling, potential role in root development
	Stress response	*miR169*	NV[b]	Induced by drought prominently in roots

References

1. Mardis, E.R., *The impact of next-generation sequencing technology on genetics.* Trends Genet, 2008. **24**(3): p. 133-41.
2. Vaucheret, H., *Post-transcriptional small RNA pathways in plants: mechanisms and regulations.* Genes Dev, 2006. **20**(7): p. 759-71.
3. Bonnet, E., Y. Van de Peer, and P. Rouze, *The small RNA world of plants.* New Phytol, 2006. **171**(3): p. 451-68.
4. Borsani, O., et al., *Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis.* Cell, 2005. **123**(7): p. 1279-91.
5. Chen, X., *Small RNAs and their roles in plant development.* Annu Rev Cell Dev Biol, 2009. **25**: p. 21-44.
6. Zimmerman, A.L. and S. Wu, *MicroRNAs, cancer and cancer stem cells.* Cancer Lett, 2011. **300**(1): p. 10-9.
7. Rayner, K.J., et al., *MiR-33 contributes to the regulation of cholesterol homeostasis.* Science, 2010. **328**(5985): p. 1570-3.
8. Meng, Y., et al., *MicroRNA-mediated signaling involved in plant root development.* Biochem Biophys Res Commun, 2010. **393**(3): p. 345-9.