Protein Cell 2015, 6(6):397–402
DOI 10.1007/s13238-015-0159-z

Protein & Cell

TOOLS

NGSmirPlant: comprehensive characterization of the small RNA transcriptomes of plants

Jie Bai[1,2], **Chen Dan**[1], **Yi Zhang**[1], **Guoping Zhao**[1,2,3]✉, **Xiaoming Ding**[1]✉

[1] Department of Microbiology and Microbial Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China
[2] Key Laboratory of Synthetic Biology, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China
[3] Shanghai-MOST Key Laboratory of Health and Disease Genomics, Chinese National Human Genome Center at Shanghai, Shanghai 201203, China
✉ Correspondence: gpzhao@sibs.ac.cn (G. Zhao), xmding74@fudan.edu.cn (X. Ding)

MicroRNAs (miRNAs) are short, endogenously expressed RNAs (~21–22 nt) processed from stem-loop regions of longer RNA precursors by a Dicer-like enzyme (Jones-Rhoades et al., 2006). In plants, microRNAs play an essential role in regulating many fundamental molecular interactions, including plant growth, development and response to environmental stress (Rogers and Chen, 2013). In comparison with the roles of microRNAs in animals, biogenesis and function of microRNAs in plants appears to be drastically different (Llave et al., 2002; Zhang et al., 2007). For example, prevalent in animal, microRNAs negatively regulate gene expression by attenuating translation. In contrast, direct microRNA-guided mRNA cleavage occurs in plant kingdom. In addition, for microRNA induced regulation in animal, translation is inhibited by microRNA with imperfect base-pairing to target mRNA. In plant, however, microRNAs inhibit the translation of target mRNA through extensively complementarity to its targets (Wu, 2013).

MicroRNAs can be identified by the approaches of hybridization-based experimental and bioinformatics prediction followed by experimental validation (Jones-Rhoades et al., 2006). With the limitations of experimental methods, microRNAs of low expression level are very difficult to be identified, causing certain microRNA to be missed. Bioinformatics approaches are often associated with a relatively high false negative and false positive rate. Thus, so far, due to limit of these methods, only a small number of plant microRNAs have been identified. Nowadays, with the advent of next-generation sequencing technologies, such as Roche 454, Illumina Solexa and ABI SOLiD, small RNA transcriptome research has been revolutionized. Compared to previously reported methods, it has distinct advantages of in-depth analysis of small RNA populations and the ability to identify novel microRNAs (Creighton et al., 2009; Fahlgren et al., 2009; Tam et al., 2014). However, deep sequencing technologies have generated hundreds of thousands of reads with large amounts of data, which give us substantial bioinformatics challenges to mine the hidden biological significance.

Currently, several useful tools have been developed to allow users to investigate the microRNAs from deep sequencing data, such as miRanalyzer (http://web.bioinformatics.cicbiogune.es/microRNA/), miRExpress (http://miRExpress.mbc.nctu.edu.tw), mirTools (http://centre.bioinformatics.zj.cn/mirtools/), miRNAkey (http://ibis.tau.ac.il/miRNAkey), DSAP (http://dsap.cgu.edu.tw), mirDeep* (http://www.australianprostatecentre.org/research/software/mirdeep-star), miREval 2.0 (http://mimirna.centenary.org.au/mireval/). However, except the UEA small RNA tool (http://srna-tools.cmp.uea.ac.uk), few tools are available for identifying plant microRNAs due to important differences of existed reference resources, mode of action and search criteria. Herein, a central motivation for this study is to specifically develop a comprehensive tool in an effort to easily investigate the large amount of NGS data for plant microRNAs. We describe a new web-based resource NGSmirPlant for the identification, characterization and differential expression analysis of microRNAs, as well as detection of novel microRNAs. Compared with the UEA small RNA tool, NGSmirPlant provides detailed annotation information for the comparison between mature microRNAs and users-input microRNAs, such as absolute counts, relative counts and sequences. In addition, NGSmirPlant implements an evolutional conservation analysis for all known microRNA families included in the result among a set of plant

---

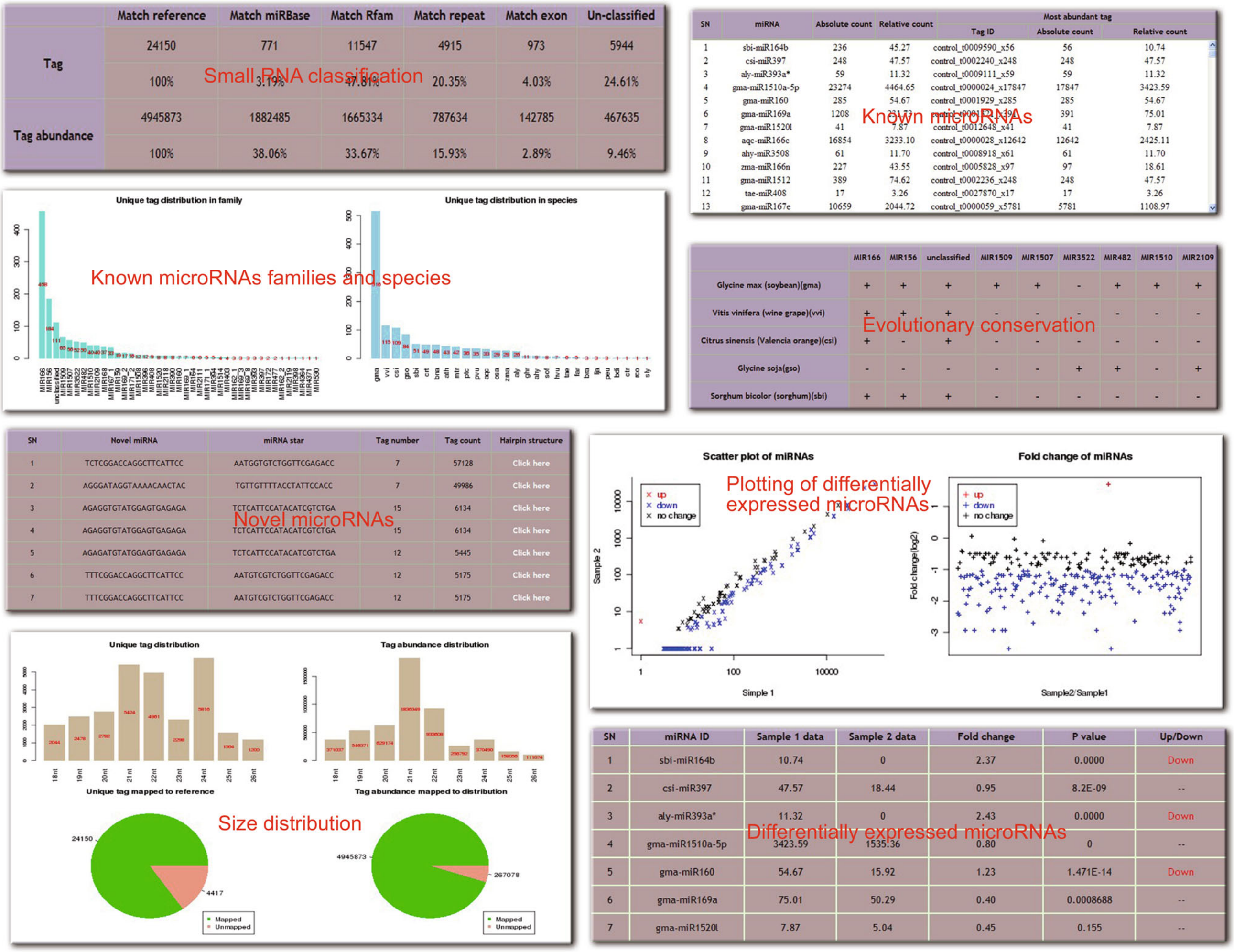
Jie Bai and Chen Dan are joint first authors.

Electronic supplementary material The online version of this article (doi:10.1007/s13238-015-0159-z) contains supplementary material, which is available to authorized users.

species. The interface of NGSmirPlant is easy to use and will be extremely helpful to plant molecular biologists with amateur bioinformatics skills to investigate the plant small RNA transcriptome.

Different small RNA has different length, which is due to distinct mechanisms of its biogenesis process. Through a read filter process, NGSmirPlant will obtain the unique length distribution of small RNA transcriptome obtained from deep sequencing, which can be used to distinguish the different classes of small RNAs and estimate the performance of sequencing library constructed. In this study, NGSmirPlant provides two kinds of length distribution (Fig. 1): one is based on unique tag and the other is based on tag abundance. The unique tag distribution is only used for estimation of the diversity of unique tag obtained from the library. The tag abundance distribution can be used to estimate the performance of sequencing library. For example,microRNAs are the most abundant class of small RNAs in plant, which will has a peak at the 21 or 22 nucleotides lengths from Dicer digestion products (Jones-Rhoades et al., 2006).

Recent studies have revealed that small RNA transcriptome is far more complex than previously thought, which might play important roles in the physiology of the cell (Creighton et al., 2009; Fahlgren et al., 2009; Tam et al., 2014). Deep sequencing has offered an exciting opportunity to determine the complexity of small RNA transcriptome in the mean of single-base resolution, genome-wide scale and cost-effectiveness. Initially, NGSmirPlant performed the



Small RNA classification

	Match reference	Match miRBase	Match Rfam	Match repeat	Match exon	Un-classified
Tag	24150	771	11547	4915	973	5944
	100%	3.19%	47.81%	20.35%	4.03%	24.61%
Tag abundance	4945873	1882485	1665334	787634	142785	467635
	100%	38.06%	33.67%	15.93%	2.89%	9.46%

Known microRNAs

SN	miRNA	Absolute count	Relative count	Most abundant tag		
				Tag ID	Absolute count	Relative count
1	sbi-miR164b	236	45.27	control_t0009590_x56	56	10.74
2	csi-miR397	248	47.57	control_t0002240_x248	248	47.57
3	aly-miR393a*	59	11.32	control_t0009111_x59	59	11.32
4	gma-miR1510a-5p	23274	4464.65	control_t0000024_x17847	17847	3423.59
5	gma-miR160	285	54.67	control_t0001929_x285	285	54.67
6	gma-miR169a	1208	231.73	[illegible]	391	75.01
7	gma-miR1520l	41	7.87	control_t0012648_x41	41	7.87
8	aqc-miR166c	16854	3233.10	control_t0000028_x12642	12642	2425.11
9	ahy-miR3508	61	11.70	control_t0008918_x61	61	11.70
10	zma-miR166n	227	43.55	control_t0005828_x97	97	18.61
11	gma-miR1512	389	74.62	control_t0002236_x248	248	47.57
12	tae-miR408	17	3.26	control_t0027870_x17	17	3.26
13	gma-miR167e	10659	2044.72	control_t0000059_x5781	5781	1108.97

Known microRNAs families and species

Unique tag distribution in family
Unique tag distribution in species

Evolutionary conservation

	MIR166	MIR156	unclassified	MIR1509	MIR1507	MIR3522	MIR482	MIR1510	MIR2109
Glycine max (soybean)(gma)	+	+	+	+	+	-	+	+	+
Vitis vinifera (wine grape)(vvi)	+	+	+	-	-	-	-	-	-
Citrus sinensis (Valencia orange)(csi)	+	-	+	-	-	-	-	-	-
Glycine soja(gso)	-	-	-	-	-	+	+	-	+
Sorghum bicolor (sorghum)(sbi)	+	+	+	-	-	-	-	-	-

Novel microRNAs

SN	Novel miRNA	miRNA star	Tag number	Tag count	Hairpin structure
1	TCTCGGACCAGGCTTCATTCC	AATGGTGTCTGGTTCGAGACC	7	57128	Click here
2	AGGGATAGGTAAAACAACTAC	TGTTGTTTTACCTATTCCACC	7	49986	Click here
3	AGAGGTGTATGGAGTGAGAGA	TCTCATTCCATACATCGTCTGA	15	6134	Click here
4	AGAGGTGTATGGAGTGAGAGA	TCTCATTCCATACATCGTCTGA	15	6134	Click here
5	AGAGATGTATGGAGTGAGAGA	TCTCATTCCATACATCGTCTGA	12	5445	Click here
6	TTTCGGACCAGGCTTCATTCC	AATGTCGTCTGGTTCGAGACC	12	5175	Click here
7	TTTCGGACCAGGCTTCATTCC	AATGTCGTCTGGTTCGAGACC	12	5175	Click here

Plotting of differentially expressed microRNAs

Size distribution

Differentially expressed microRNAs

SN	miRNA ID	Sample 1 data	Sample 2 data	Fold change	P value	Up/Down
1	sbi-miR164b	10.74	0	2.37	0.0000	Down
2	csi-miR397	47.57	18.44	0.95	8.2E-09	--
3	aly-miR393a*	11.32	0	2.43	0.0000	Down
4	gma-miR1510a-5p	3423.59	1535.36	0.80	0	--
5	gma-miR160	54.67	15.92	1.23	1.471E-14	Down
6	gma-miR169a	75.01	50.29	0.40	0.0008688	--
7	gma-miR1520l	7.87	5.04	0.45	0.155	--

Figure 1. The typical output of NGSmirPlant web server. It includes: 1) the size distribution of small RNA transcriptome; 2) the classification of small RNAs known microRNAs, repeat-associated RNA, the ncRNAs reads annotated by Rfam (rRNA, tRNA, snRNAs and snoRNA et al.) and mRNA degradation products; 3) the known and novel microRNAs identification; 4) evolutionary conservation of individual microRNA in various plant species; and 5) differentially expressed microRNAs between samples.

mapping of sequence tags mapping to the reference genome or EST using SOAP 2.0 (http://miRExpress.mbc.nctu.edu.tw). It is found that, besides high abundance of microRNA, non-coding small RNAs (snRNAs, snoRNAs and tRNAs et al. obtained from the Rfam database) will also occupy many sequence tags even for the specific designed microRNA sequencing library (http://122.228.158.106/NGSmirPlant/download1.php?id=1414376897). In addition, a small proportion of sequencing tags could be mapped to exon sequences, which are likely to be RNA degradation products. Therefore, the aligned reads are classified into known microRNAs, repeat-associated RNA, the ncRNAs reads annotated by Rfam (rRNA, tRNA, snRNAs and snoRNA et al.) and mRNA degradation products (Fig. 1). If a sequence tag has multiple hits, we only assign it a single category based on the following priority: ncRNA from Rfam, known plant microRNA, repeat-associated RNA and mRNA degradation product. Through such a way, NGSmirPlant will identify the expression level of each kind of small RNA by calculating the number of total tags it belongs to.

One of the most important advantages for deep sequencing of small RNA transcriptome is that it can be used for the discovery of novel microRNAs (Creighton et al., 2009). In NGSmirPlant, to identify candidate novel microRNAs, we first filter out sequence tags that have been classified into annotated categories, such as non-coding RNA and known microRNAs. With the implement of the miRDeep program (Friedlander et al., 2008), NGSmirPlant provides candidate novel microRNAs with the RNA secondary structure, minimal free energy, tag counts of mature/mature* sequence et al. (Fig. 1). Users can use other method, such as hybridization-based experimental method, to validate the reliability of novel microRNA identified.

Since many microRNA families are evolutionarily conserved across many lineages in the plant kingdom and the fact that lots of microRNAs in different plant species shared the same mature sequence in miRBase 19.0 (http://microrna.sanger.ac.uk/), microRNA genes in one species may have orthologs or homologs existed in other species. Given that, we enlarged the annotation scale to an expansion of multiple species of known and novel microRNAs. For the identified microRNAs, the measurement of its evolutionary conservation was performed by similarity search against the currently known microRNAs in miRBase 19.0 in other plant species (Fig. 1). If a read matched one or more known microRNAs with >95% identity, it was regarded as homologous in evolution. Through the evolutionary conservation profile, the lineage-specific and species-specific microRNAs can be identified and visualized easily.

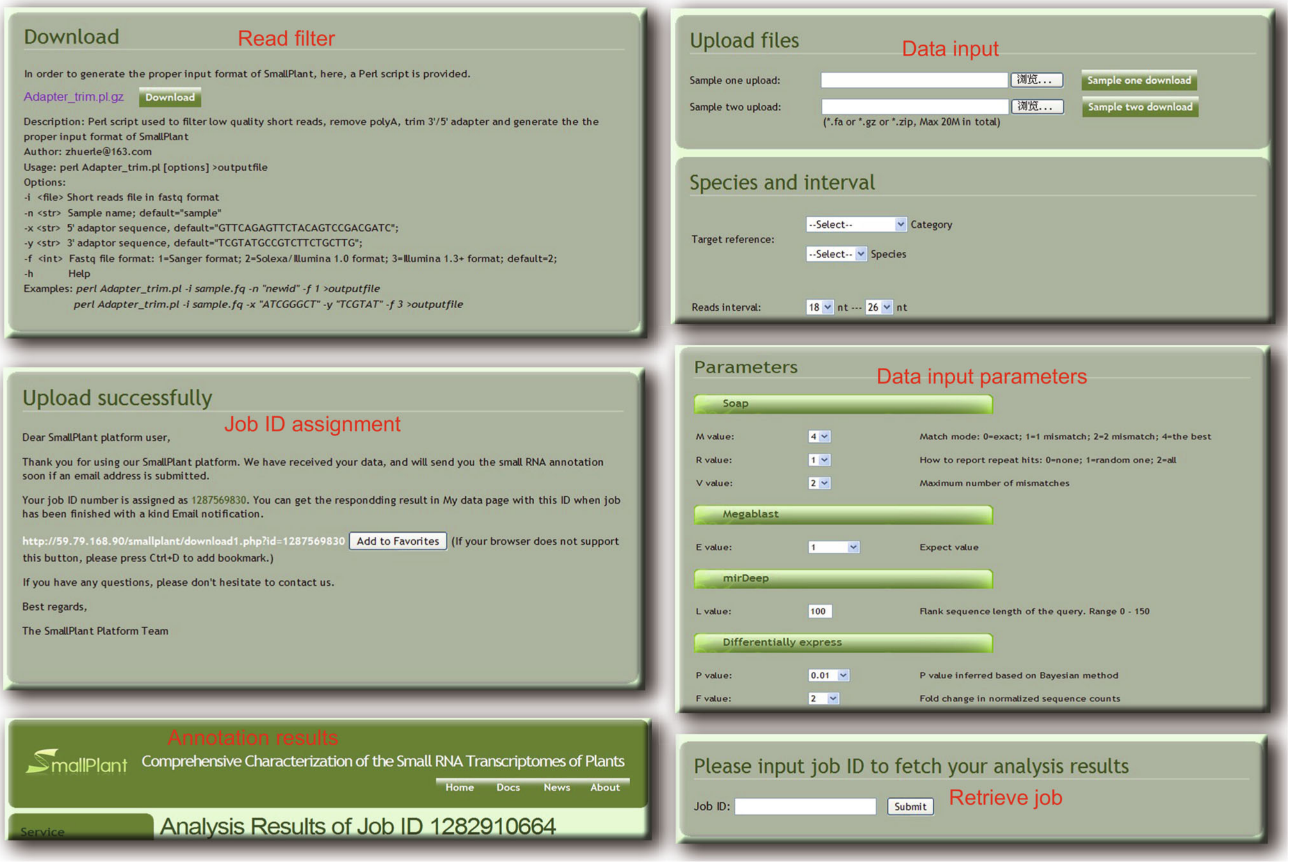
NGSmirPlant employs two different strategies to investigate the microRNAs expression profile and detect the differentially expressed microRNAs between different libraries (Fig. 1). One is based on the total tag count (#specific microRNA tags/#total microRNA sequence tags) and the other is based on the most abundant tag count (#the most abundant tag of specific microRNA/#total microRNA sequence tags). Each identified microRNA sequence numbers were normalized to the total number of microRNA read counts followed by multiplying one million to obtain reads per million (RPM) value. The statistical significance (*P*-value) is inferred based on the Bayesian method (Audic and Claverie, 1997). Flexibly, the threshold for fold change as differentially expressed microRNA is user-definable. NGSmirPlant provides differentially expressed microRNAs with microRNA ID, relative sequenced count, fold change, increase/decrease and *P* value.

The web server of NGSmirPlant is well organized in a user-friendly access manner. It runs with the input of a single sample or multiple samples. For “single” mode, NGSmirPlant will analyze and annotate small RNAs embed in the sample and give a list of novel microRNA details. For “multiple” mode, the differentially expressed microRNAs between the pair samples will be in-depth investigated additionally.

The web server input of NGSmirPlant is pre-processed FASTA format where all the unique sequence tags aggregated with a number behind the name to represent the expression level and cleaned into a non-redundant FASTA file to reduce the input size. The users can download the Perl script in the filter column of web server to a local client for generating the proper input format of NGSmirPlant. The size of input file should be limited to 10 MB, which can be in FASTA format or compressed in zip or gz format of FASTA file (Fig. 2). A demo for the web server is provided. Users can download the example input file for “single” or “multiple” mode and run the tools with default parameters. In general, a typical “single” run takes about 20 min and a typical “multiple” run takes about 40 min.

NGSmirPlant provide a number of important parameters in both of “single” mode and “multiple” mode, so users can set these parameters by themselves (Fig. 2). Users can set the reads length interval (default is 18~30 nt) thus the tag sequences longer or shorter than the threshold will be discarded. During the alignment mapping process, a set of parameters are available for users to define. With the “R value” parameter, users can freely define the strategy for reporting repeat hits. It should be noticed that the “M value” parameter denotes the maximum number of allowed mismatches for the seed part of a read (at most two mismatches) while the “V value” parameter means the total allowed mismatches in one read. When predicting novel microRNAs, the expected value and the flanking sequence length of the query can be defined by users. For “multiple” mode, two parameters for differentially expression are provided. By setting the threshold of *P* value inferred based on Bayesian method and fold change in normalized sequence counts, the specific microRNA satisfied both the *P*-value and the fold change will be deemed to be significantly differentially expressed in the two samples (Fig. 3).

Once a job is submitted successfully, NGSmirPlant will be assigned with a unique job ID and users need to wait for the result by e-mail provided. Then, users will receive a notification e-mail with a URL link automatically. Users can use



Download

In order to generate the proper input format of SmallPlant, here, a Perl script is provided.

Adapter_trim.pl.gz

```
Description: Perl script used to filter low quality short reads, remove polyA, trim 3'/5' adapter and generate the the
proper input format of SmallPlant
Author: zhuerle@163.com
Usage: perl Adapter_trim.pl [options] >outputfile
Options:
-i  <file> Short reads file in fastq format
-n <str>  Sample name; default="sample"
-x <str>  5' adaptor sequence, default="GTTCAGAGTTCTACAGTCCGACGATC";
-y <str>  3' adaptor sequence, default="TCGTATGCCGTCTTCTGCTTG";
-f  <int>  Fastq file format: 1=Sanger format; 2=Solexa/Illumina 1.0 format; 3=Illumina 1.3+ format; default=2;
-h         Help
Examples: perl Adapter_trim.pl -i sample.fq -n "newid" -f 1 >outputfile
          perl Adapter_trim.pl -i sample.fq -x "ATCGGGCT" -y "TCGTAT" -f 3 >outputfile
```

Upload successfully

Dear SmallPlant platform user,

Thank you for using our SmallPlant platform. We have received your data, and will send you the small RNA annotation soon if an email address is submitted.

Your job ID number is assigned as 1287569830. You can get the respondding result in My data page with this ID when job has been finished with a kind Email notification.

http://59.79.168.90/smallplant/download1.php?id=1287569830 (If your browser does not support this button, please press Ctrl+D to add bookmark.)

If you have any questions, please don't hesitate to contact us.

Best regards,

The SmallPlant Platform Team

SmallPlant Comprehensive Characterization of the Small RNA Transcriptomes of Plants

Analysis Results of Job ID 1282910664

Upload files

Sample one upload:

Sample two upload:

(*.fa or *.gz or *.zip, Max 20M in total)

Species and interval

Target reference: --Select-- Category; --Select-- Species

Reads interval: 18 nt --- 26 nt

Parameters

M value:	4	Match mode: 0=exact; 1=1 mismatch; 2=2 mismatch; 4=the best
R value:	1	How to report repeat hits: 0=none; 1=random one; 2=all
V value:	2	Maximum number of mismatches
E value:	1	Expect value
L value:	100	Flank sequence length of the query. Range 0 - 150
P value:	0.01	P value inferred based on Bayesian method
F value:	2	Fold change in normalized sequence counts

Please input job ID to fetch your analysis results

Job ID:

Figure 2. The typical input of NGSmirPlant showing the reads filter requirement, file input, job ID assignment and job retirement.

this unique ID to visit the annotation results or retrieve previous job, which will be kept one month in NGSmirPlant. The job ID submitted older than one month will be deleted. Subsequently, NGSmirPlant will show: 1) basic summary information of library, such the length distribution of unique sequence size, which can be used to indicate the most abundant of small RNA category in the library; 2) alignment information indicating the tag abundance matched to reference genome, miRBase, Rfam, repeat, exon and unannotation respectively; 3) detailed annotation information of the known microRNA identified, such species and families distribution, the absolute (total number of sequence tags mapped to a specific microRNA) and relative (normalization of matched read counts to the total number of microRNA reads and then multiplied by a million) expression level and evolutionary conservation of individual microRNA in various plant species; 4) detailed annotation information of other non-coding RNA identified, such as rRNA, tRNA, snRNAs and snoRNA; 5) detailed annotation information of novel microRNA identified, such as novel microRNA sequence, microRNA star and hairpin structure; 6) detailed information of microRNA different expression. All these results can be downloaded from NGSmirPlant to allow users to perform easy visualization or further analysis in the local computer.

The small RNAs, especially the microRNAs, play important roles in different biological processes in plant, such as plant growth, development and response to environmental stress. However, currently, only a small number of plant microRNAs have been detected due to the limitation of hybridization-based experimental and bioinformatics prediction methods. However, the recently developed deep sequencing technologies have been proved to have distinct advantages of unbiased and in-depth analysis of the complexity of small RNA transcriptome and its powerful ability to identify known and novel microRNAs. However, next-generating sequencing technologies have produced hundreds of thousands of short reads with large amounts of data, which give us substantial bioinformatics challenges to discover the hidden biological significance. Therefore, in this study, NGSmirPlant (http://122.228.158.106/NGSmirPlant/) is specifically designed towards a web-based server to fulfill the urgent need for bioinformatics analysis of plant small RNA transcriptome by deep sequencing. In general, NGSmirPlant can be used to facilitate a genome-wide analysis of microRNAs profiling and discovery, exploring evolutionary conservation of individual microRNA in various plant species.

In comparison with other microRNA profiling tools, such as mirTools (http://centre.bioinformatics.zj.cn/mirtools/) and

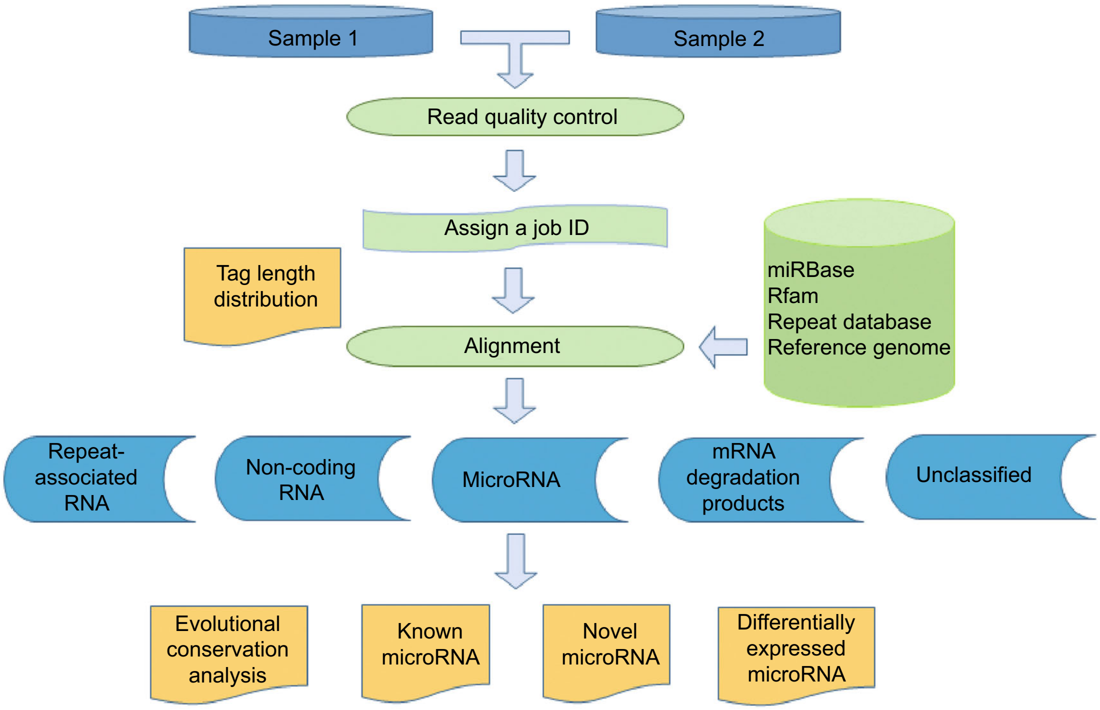


Figure 3. The workflow of NGSmirPlant. In brief, NGSmirPlant firstly performs the read quality control and aligns the clean reads to a number of reference databases, such as genomic sequences or ESTs, miRBase, Rfam and repeat database produced by RepeatMasker. The small RNA annotation categories will be then identified, include microRNAs, repeat-associated RNA, non-coding RNA and mRNA degradation products. Meanwhile, NGSmirPlant will provide detailed annotation of known and novel microRNAs, differentially expressed microRNA between samples and its evolutional conservation.

CPSS (http://mcg.ustc.edu.cn/db/cpss/index.html) which are either specialized for animal small RNAs analysis or incomplete analysis of plant small RNAs, NGSmirPlant is specialized for comprehensive characterization of the small RNA transcriptome of plants. Currently, it covers a broad range of 85 plant organisms from the KEGG database (http://www.genomejp/kegg/catalog/org_list.html, includes completely finished genomic sequences and the available ESTs) to allow users to effectively identify known and novel microRNAs. It is believed that the comprehensive integration of plant species could meet the need of most researchers who involve in plant small RNA analysis from deep sequencing. Moreover, NGSmirPlant performs evolutionary conservation analysis for each microRNA family while mirTools and CPSS do not. In this part, those relatively conserved microRNA families among different species are listed with a symbol "+". In addition, the detailed annotation of known microRNAs is performed for the comparison between mature microRNAs and users-input microRNAs, such as absolute counts, relative counts and sequences. Compared with CPSS, NGSmirPlant exhibits more readable graphs, such as the detailed distribution of unique tags in families and species, which provides a more concise result for users to visualize the results. As with other microRNA profiling tools, such as the UEA small RNA tool, NGSmirPlant also perform an analysis of differentially expressed microRNAs between samples. Meanwhile, it is very useful that NGSmirPlant draws scatter plots for total tag count and the most abundant tag, which makes the expression differentiation extremely clear.

In the future version, we will continually integrate more number of finished plant genomes or ESTs data as they become increasingly available. In addition, the pre-processing script provided in the current version only supports with Illumina sequencing data and as such would require users to develop their own pre-processing steps in order to use 454 or SOLiD data. In the future, a more general and flexible pre-processing script will be made available that would accept the raw data obtained from the three common used high-throughput sequencing platforms and generate the required format for NGSmirPlant input. To further improve on the power of NGSmirPlant, future development will implement the microRNA target prediction tools to predict the target genes of identified microRNAs in plant. Moreover, in addition to the microRNA target prediction tools, NGSmir-Plant will plan to perform detailed functional annotation of microRNA targeted genes, including Gene Ontology, KEGG pathway and protein-protein interaction. Through which, NGSmirPlant will provide the opportunities to elucidate the complex microRNAs mediated regulatory networks.

NGSmirPlant welcomes any questions, comments and suggestions, which will be a useful feedback for future updating. In conclusion, we believe that NGSmirPlant will be a straight-forward valuable resource that are opening fascinating opportunities for in-depth investigation of small RNA transcriptome generated from next-generation sequencing technologies.

NOTES

This work was supported by National Basic Research Program (973 Program) (No. 2012CB721102), National Natural Science Foundation of China (Grant No. 30830002) and the China Postdoctoral Science Foundation funded project (2012T50444 and 2012M 520947).

Jie Bai, Chen Dan, Yi Zhang, Guoping Zhao and Xiaoming Ding declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by the any of the authors.

OPEN ACCESS

REFERENCES

Audic S, Claverie JM (1997) The significance of digital gene expression profiles. Genome Res 7:986–995

Creighton CJ, Reid JG, Gunaratne PH (2009) Expression profiling of microRNAs by deep sequencing. Brief Bioinform 10:490–497

Fahlgren N, Sullivan CM, Kasschau KD, Chapman EJ, Cumbie JS, Montgomery TA, Gilbert SD, Dasenko M, Backman TW, Givan SA, Carrington JC (2009) Computational and analytical framework for small RNA profiling by high-throughput sequencing. Rna 15:992–1002

Friedlander MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, Rajewsky N (2008) Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol 26:407–415

Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAS and their regulatory roles in plants. Annu Rev Plant Biol 57:19–53

Llave C, Kasschau KD, Rector MA, Carrington JC (2002) Endogenous and silencing-associated small RNAs in plants. Plant Cell 14:1605–1619

Rogers K, Chen X (2013) Biogenesis, turnover, and mode of action of plant microRNAs. Plant Cell 25:2383–2399

Tam S, de Borja R, Tsao MS, McPherson JD (2014) Robust global microRNA expression profiling using next-generation sequencing technologies. Lab Investig 94:350–358

Wu G (2013) Plant microRNAs and development. J Genet Genomics 40:217–230

Zhang B, Wang Q, Pan X (2007) MicroRNAs and their regulatory roles in animals and plants. J Cell Physiol 210:279–289