View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by International Journal for Computational Biology (IJCB)

International Journal for Computational Biology (IJCB)
Vol.5, No.1, August 2016, pp. 21~27
ISSN: 2278-8115

Development of a New Pipeline for Identification and Characterization of Micro RNAs from Plants

Jinu Thomas and K K Sabu*
Division of Biotechnology and Bioinformatics, Jawaharlal Nehru Tropical Botanical Garden and Research Institute, Palode, Thiruvananthapuram 695562, Kerala, India

Article Info

Article history:

Received Jul 26th, 2015
Revised Aug 20th, 2015
Accepted Sep 24th, 2015

Keyword:

MicroRNA
miRNA
Analysis pipeline
NGS
In silico analysis
Next generation sequencing

ABSTRACT

Open source microRNA analysis pipelines for next generation sequencing data (NGS) often make necessary use and working knowledge of command line interface, massive data processing resources and expertise which is a daunting task for biologists. Further, the microRNA data generated from NGS platforms will not be in a form from which one could understand or make use of it. Hence a comprehensive pipeline has been developed by integrating several open source NGS tools along with a graphical user interface called sRNAbench. It is useful for expression profiling of small RNAs and prediction of microRNAs from NGS data. The pipeline features functionalities such as read processing, sequence identification, target prediction and enrichment analysis. It provides even prediction of novel microRNAs and its sequences. The pipeline will be very useful for plant genomics community and it does not require knowledge in computational biology in order to discover miRNAs and utilize the same in genomics studies.

Copyright © 201X ***International Journal for Computational Biology****,* ***http:// www.ijcb.in****, All rights reserved.*

Corresponding Author:

K K Sabu,
Jawaharlal Nehru Tropical Botanical Garden and Research Institute, India.
Email: sabu@jntbgri.res.in

How to Cite:

Thomas *et. al.*, Development of a New Pipeline for Identification and Characterization of Micro RNAs from Plants. IJCB. 2016; Volume 5 (Issue 1): Page 21-27.

1. INTRODUCTION

The central dogma of molecular biology, *i.e.* the flow of biological information from DNA to RNA to protein, has been challenged by the discovery of a gene regulation mechanism, known as RNA-mediate gene silencing (RNA silencing). It is the process whereby small, approximately 20-24 nucleotide (nt), non-coding RNA molecules direct the sequence-specific down-regulation of target genes at the post-transcriptional level. There are two known classes of RNAs to mediate this process, namely microRNAs (miRNAs) and short interfering RNAs (siRNAs), which act as cellular regulators to control mRNA stability and translation, to protect the genome from the invading nucleic acids, to facilitate epigenetic modifications of chromatin and histones, and to direct complex developmental pathways [1].

The term ‘microRNA’ was first coined in 2001 when tens of small RNAs with regulatory potential were discovered in *Caenorhabditis elegans* [2]. Later, the scientific world witnessed large number of research programmes in this area. For example, a term search for ‘microRNAs’ in PubMed clearly reveals the increasing interest by listing about 36,740 research papers as on July 2015. These research were made possible through next-generation sequencing (NGS) which allows the sequencing of small RNA molecules and the estimation of their expression levels. Consequently, there is high demand of bioinformatics tools to cope with the several gigabytes of sequence data generated in each of the deep-sequencing experiments.

Plant miRNAs target recognition mechanism was once thought to be simple and straightforward. Therefore, very few target prediction tools and algorithms were developed for plants as compared to those for animals. However, later studies revealed the enormous diversity and complexity of the gene regulation by miRNAs in plant systems. This, in turn, necessitates the need for advanced computational tools/algorithms for comprehensive miRNA target analysis to help understand miRNA regulatory mechanisms [3]. In this regard, many new programs such as psRNATarget was developed [4] for target prediction of plant endogenous non-coding short small RNAs. miRBase was created as an online repository for miRNA sequence data which was integrated with interfaces for comprehensive analysis of miRNA sequence data, annotation and prediction of gene targets [5]. Later, another powerful platform named miRanalyzer: a miRNA detection and analysis tool for NGS experiments was launched [6,7]. Subsequently, in order to find differentially expressed miRNAs and to predict novel miRNA candidates, miRSeqNovel was developed [8].

Even though excellent tools are available for miRNA analysis, no pipeline has been set-up using tools available in the public domain which require minimum computational knowledge. Hence, the present study was aimed at developing an analysis pipeline for miRNA sequence data using those freely available tools. The pipeline was formulated by analyzing the miRNAs that are expressed in *Brachypodium distachyon* roots and leaves under drought stress conditions (details of the experiment was available at NCBI database under SRA id 160390).

2. RESEARCH METHOD

2.1. Sequence retrieval from SRA

RNA sequence reads of *Brachypodium distachyon* leaves and roots in control and drought stress conditions were accessed from NCBI SRA database (SRA 160390).

2.2. Read processing using sRNAbench

The reads were processed using sRNAbench (http://bioinfo5.ugr.es/srnatoolbox/srnabench), replacement for popular miRanalyzer program. It is a web based tool for processing small RNA reads that are obtained from high throughput sequencing data. The datasets were provided as input to the program by means of URL obtained from the SRA database. Then the 'Do not map to genome (Library mode)' check box was activated followed by selecting *Arabidopsis thaliana* (tair 10) as a reference species (*Brachypodium distachyon* was not listed in the sRNAbenc as a reference species during our analysis). The sRNAbench will use the annotations from the sRNAbench database for the selected species during the reads mapping. If no species is selected and the 'Do not map to genome (Library mode)' check box is not activated, then, sRNAbench will only analyse microRNAs. If a species is selected and the 'Do not map to genome (Library mode)' check box is not activated sRNAbench will use the genome mapping mode. Adapter trimming was done by selecting Illumina RA3 with Recursive Adapter trimming selected. Then, the microRNA analysis was done for the species selected during the Select species step. Then the analysis parameters were specified following minimum read count as 2, number of allowed mismatches as 0, seed length of alignment as 20, alignment type as 'n', remove barcode as 0, minimum read length to predict new miRNAs as 15 and maximum number of multiple mapping as 10. The sRNAbench job IDs were used to specify the differential expression analysis.

2.3 microRNA sequence identification using miRBase

miRBase (http://www.mirbase.org) is a biological database that acts as an archive of miRNA sequences and annotations. The miRBase registry provides a centralized system for assigning new names to miRNA genes. The sequences obtained from the sRNAbench were given as input to miRBase for sequence identification.

2.4. microRNA target prediction using psRNATarget

psRNATarget, is a plant small RNA target analysis server, which features two important analysis functions: reverse complementary matching between miRNA and target transcript using a proven scoring schema, and target site accessibility evaluation by calculating unpaired energy (UPE) required to "open" secondary structure around miRNA's target site on mRNA. The sequence obtained from miRBase was given as input to the psRNATarget and selected preloaded transcript or genomic library from the available list. This resulted in a list of predicted miRNA/Target pairs with its alignment.

Table 1. Details of the reads, read quality, mapping and miRNAs obtained from the *Brachypodium distachyon* 3rd leaf division zone control and drought stress experiment as obtained from the new pipeline

Experiment	*B. distachyon* (Bd21) 3rd leaf division zone			
	Control sequencing experiment 1	Control sequencing experiment 2	Drought stress sequencing experiment 1	Drought stress sequencing experiment 2
SRA id	SRR522515	SRR522516	SRR522517	SRR522518
sRNAbench job id	57728910	7873484	29856730	53623643
Raw reads	29011509	28846684	19783226	26287050
Adapter trimmed	22294230 (76.8%)	22652738 (78.5%)	16375137 (82.8%)	25647475 (97.6%)
Length filtered reads	299966 (1.0%)	2500425 (8.7%)	410885 (2.1%)	638367 (2.4%)
Quality filtered reads	141088 (0.5%)	106517 (0.4%)	56918 (0.3%)	136776 (0.5%)
Reads in analysis	21853176 (75.3%)	20045796 (69.5%)	15907334 (80.4%)	24872332 (94.6%)
Genome mapped reads	2208101 (10.10%)	5671628 (28.29%)	1796065 (11.29%)	1945515 (7.82%)
Detected mature miR	45 (10.54%)	42 (9.84%)	45 (10.54%)	36 (8.43%)
Reads mapped to miRBase hairpins	1498399	316449	1133591	842891
isomiR length variants	67.86 %	5.58 %	63. 12 %	43.32 %

3.5. Enrichment analysis using agriGO and REVIGO

The results obtained from the psRNATarget was given to agriGO toolkit as in the format of gene list obtained from phytozome which enabled analysis of GO relationships in graphical format along with GO terms, GO source and its description. This obtained separate graphs for each of the three GO categories, namely biological process, molecular function and cellular component. Singular enrichment analysis was carried out as to obtain desired results. REVIGO was used to create semantic similarity based scatter plots by removing redundancy which makes the results more easier to interpret.

3. RESULTS AND ANALYSIS

The role played by miRNAs at various stages of development was worked out in many plant species. For example, the role of miRNAs in drought response was investigated in young leaves of *Brachypodium distachyon*, a drought-tolerant monocot model species. NGS data analysis identified 66 annotated miRNA genes and 122 new high confidence predictions greatly expanding the number of known *Brachypodium* miRNAs. Most miRNAs showed a high expression level, consistent with their involvement in early leaf development and cell identity [9].

In the present study, the raw data from the 8 reads (SRX160390 though SRX160397) deposited at the NCBI SRA database after NGS were processed using sRNAbench. Tabulated outputs of the sRNAbench analysis are given as Table 1 summarizing read quality, genome mapped reads and MIR profiling results. Using sRNAbench our read was mapped to *Arabidopsis thaliana* (tair 10). Total of 338 mature miRNAs were detected

and one novel miRNA was predicted with a total read count of 2017. The differential expression analysis using sRNAbench provided list of miRNAs that are differentially expressed. The miRNA sequences for the differentially expressed ones were fetched using miRBase (Table 2). Subsequently, miRNA target prediction was carried out using psRNA Target.

Table 2. miRNA sequence obtained from miRBase based on the result inferred from sRNAbench analysis

miRNA	Sequence
at-miR169a-5p	>ath-miR169a-5p MIMAT0000200 CAGCCAAGGAUGACUUGCCGA
ath-miR396a-3p	>ath-miR396a-3p MIMAT0031908 GUUCAAUAAAGCUGUGGGAAG
ath-miR8175	>ath-MIR8175 MI0026805 CUUAAGGCAAUUUGACCUAUAUCAAAUUUGGCACCGUUGCCGGGGAUCGAC CGGGUCACCCGCGUGACAGGCGGGAAUCUUACCACUUAGUACAACGACCCA AUAUAGUGGUAAGUAUUCUCGCAUGUCACCGGGUUCGAUCCCCGGCAACGG CGCCCAUUUGAUAUAGCUCAAAUCGCCUUAA
ath-miR167d	>ath-MIR167d MI0000975 UGUUGGUUUUUAGAAGCUGAAGCUGCCAUGAUCUGGUAAUCGCUACAUACG ACACACACAUCACUAAACUUCUUUAUAAUAUGCACACACAUACAGCUCUUA AUGGCAACUCAAAGUUAUAAUUAGUGCAUGACCUAGUUAUUUGACUGCUU UUAAUAUAUUUAUGGAUUCACGCAUGUGUGUGUAUGCAUAAUUUACAUGC AUGCACUUUGUGUGUACACAUCAAUUUGAACCCGUUCAAAUCUGUUUUUAU UAGUAUAUAUAUAGAUGUGUGGUGUGUGUGUCAGUGUGUGUGUGUUUAUA GAUAGUAGUACUAGGUCAUCGCAGCUUCAGUCACUAAAUCACCAACA
ath-miR160a-5p	>ath-miR160a-5p MIMAT0000178 UGCCUGGCUCCCUGUAUGCCA
ath-miR160b	>ath-MIR160b MI0000191 GUCGUGCCUGGCUCCCUGUAUGCCACAAGAAACAUCGAUUUAGUUUCAAAA UCGAUCACUAGUGGCGUACAGAGUAGUCAAGCAUGA
ath-miR160c-5p	>ath-miR160c-5p MIMAT0000180 UGCCUGGCUCCCUGUAUGCCA
ath-miR171b-3p	>ath-miR171b-3p MIMAT0000920 UUGAGCCGUGCCAAUAUCACG
ath-miR171c-3p	>ath-miR171c-3p MIMAT0000921 UUGAGCCGUGCCAAUAUCACG
ath-miR164c-5p	>ath-miR164c-5p MIMAT0001017 UGGAGAAGCAGGGCACGUGCG
ath-miR172c	>ath-MIR172c MI0000991 AGCUACUGUUCGCUGUUGGAGCAUCAUCAAGACACAAAUCAUCAAGUAUUC GUGUAAAUAAUUACCAUUUAUGAUUAGAUUUUUGAUGUAUGUAUGAAUCU UGAUGAUGCUGCAGCUGCAAUCAGUG
ath-miR172d-3p	>ath-miR172d-3p MIMAT0000923 AGAAUCUUGAUGAUGCUGCAG
ath-miR319b	>ath-MIR319a MI0000544 AGAGAGAGCUUCCUUGAGUCCAUUCACAGGUCGAUUCAAUUAGCUUCCGAC UCAUUCAUCCUAUAAUACCGAGUCGCCAAAAUUCAAACUAGACUCUAAAUG AAUGAAUGAUGCGGUAGACAAAUUGCAUUGAUUCUCUUUGAUUGGACUGA AGGGAG
ath-miR396a-5p	>ath-miR396a-5p MIMAT0000944 UUCCACAGCUUUCUUGAACUG
ath-miR396b-3p	>ath-miR396b-3p MIMAT0031909 GCUCAAGAAAGCUGUGGGAAA
ath-miR156j	>ath-MIR156j MI0019234 UUUGACAGAAGAGAGAGAGCACAGUUUGAG CAAUUCCCUUGUAGCUCUCUAGUUUAGUGUC
	>ath-MIR156g MI0001082

ath-miR156g	AUAACGAAGGCGACAGAAGAGAGUGAGCACAUUUUUCUAGCAUGCUCAUGC UCGAAAGCUCUCUUACUCUCUUCUUGUCUCCUGCUCUCU
ath-miR159b-3p	>ath-miR159b-3p MIMAT0000207 UUUGGAUUGAAGGGAGCUCUU
ath-miR159a	>ath-MIR159a MI0000189 GUAGAGCUCCUUAAAGUUCAAACAUGAGUUGAGCAGGGUAAAGAAAAGCU GCUAAGCUAUGGAUCCCAUAAGCCCUAAUCCUUGUAAAGUAAAAAAGGAUU UGGUUAUAUGGAUUAUAUCUCAGGAGCUUUAACUUGCCCUUUAAUGGCUU UUACUCUUCUUUGGAUUGAAGGGAGCUCUAC
ath-miR393a-5p	>ath-miR393a-5p MIMAT0000934 UCCAAAGGGAUCGCAUUGAUCC
ath-miR393b-5p	>ath-miR393b-5p MIMAT0000935 UCCAAAGGGAUCGCAUUGAUCC
ath-miR408-3p	>ath-miR408-3p MIMAT0001011 AUGCACUGCCUCUUCCCUGGC
ath-miR172e-3p	>ath-miR172e-3p MIMAT0001019 GGAAUCUUGAUGAUGCUGCAU
ath-miR156c-3p	>ath-miR156c-3p MIMAT0031867 GCUCACUGCUCUAUCUGUCAGA
ath-miR395a	>ath-MIR395a MI0001007 AUGUCUCCUAGAGUUCCUCUGAGCACUUCAUUGGGGAUACAAUUUUUCUAA AUGAUUAUCCACUGAAGUGUUUGGGGGAACUCCCGGACCCAU
ath-miR395d	>ath-MIR395d MI0001010 AUGUCCUCUAGAGUUCUCCUGAACACUUCAUUGGAAAUUUGUUAUUCAGUA AGCUAACAGUUAAUUCCACUGAAGUGUUUGGGGGAACUCCCGAUG
ath-miR395e	>ath-MIR395e MI0001011 AUGUUUUCUAGAGUUCCUCUGAGCACUUCAUUGGAGAUACAAUUUUUUAU AAAAUAGUUUUCUACUGAAGUGUUUGGGGGAACUCCCGGGCUGAU
ath-miR166a-3p	>ath-miR166a-3p MIMAT0000189 UCGGACCAGGCUUCAUUCCCC
ath-miR166b-3p	>ath-miR166b-3p MIMAT0000190 UCGGACCAGGCUUCAUUCCCC
ath-miR166	>ath-MIR166c MI0000203 GCGAUUUAGUGUUGAGAGGAUUGUUGUCUGGCUCGAGGUCAUGAAGAAGA GAAUCACUCGAAUUAAUUGGAAGAACAAAUUAAGAAAACCCUAGAUGAUU CUGGACCAGGCUUCAUUCCCCCUAACCUACUUAUCGC

From the miRNA target analysis, 118 target genes were obtained for the differentially expressed 32 miRNA sequences, from which gene list was prepared (Bradi1g70720.1, Bradi3g28970.1, Bradi1g14940.1, Bradi1g53650.4, Bradi1g50597.1, Bradi1g36540.1, Bradi4g01887.1, Bradi3g32890.1, Bradi2g37800.2, Bradi3g52547.1, Bradi2g53010.1, Bradi4g01887.1, Bradi1g52240.1, Bradi2g37800.1, Bradi3g52547.2, Bradi3g52980.1, Bradi5g18830.1, Bradi1g78230.1, Bradi1g30337.1, Bradi5g20607.1, Bradi1g60120.1, Bradi1g01640.1, Bradi1g08847.1, Bradi5g24100.1, Bradi4g16450.2, Bradi3g28950.1, Bradi4g42720.3, Bradi1g61130.1, Bradi1g38715.1, Bradi1g12650.3, Bradi3g06487.1, Bradi4g42720.1, Bradi1g10780.1, Bradi4g16486.1, Bradi3g57267.1, Bradi3g28950.1, Bradi1g47670.1, Bradi1g10780.2, Bradi2g31250.1, Bradi4g16450.1, Bradi3g06487.1, Bradi1g13910.1, Bradi3g32890.1, Bradi2g31250.2, Bradi1g09900.1, Bradi4g02060.1, Bradi2g06210.1, Bradi1g52240.1, Bradi5g23120.1, Bradi1g52150.2, Bradi1g32660.1, Bradi3g28970.1, Bradi1g78230.1, Bradi3g51950.3, Bradi4g05940.1, Bradi1g41710.1, Bradi5g18830.1, Bradi1g08847.1, Bradi3g51950.1, Bradi1g12650.1, Bradi1g03207.1, Bradi3g51590.1, Bradi1g61130.1, Bradi3g34737.1, Bradi1g52150.3, Bradi3g58990.2, Bradi4g01887.1, Bradi1g10780.1, Bradi3g34737.2, Bradi4g27220.1, Bradi3g58990.1, Bradi1g47670.1, Bradi1g10780.2, Bradi5g08680.1, Bradi2g14990.1, Bradi2g59640.1, Bradi1g13910.1, Bradi1g03880.1, Bradi2g35720.1, Bradi1g18200.1, Bradi1g62140.1, Bradi2g06210.1, Bradi1g53650.1, Bradi5g08680.1, Bradi1g03530.2, Bradi1g62140.2, Bradi3g28970.1, Bradi1g53650.4, Bradi2g35720.1, Bradi1g03180.1, Bradi4g42720.2, Bradi4g01887.1, Bradi2g37800.2, Bradi3g22020.4, Bradi3g27912.1, Bradi4g42720.3, Bradi3g51590.1, Bradi2g37800.1, Bradi3g22020.1, Bradi3g39630.1, Bradi4g42720.1, Bradi5g18830.1, Bradi1g30337.1, Bradi3g22020.3, Bradi1g21700.1, Bradi1g47670.1, Bradi1g01640.1, Bradi4g39540.1, Bradi3g22020.6, Bradi3g54890.1, Bradi1g13910.1, Bradi2g58570.1, Bradi1g12540.1, Bradi1g03610.1, and Bradi4g21120.1).

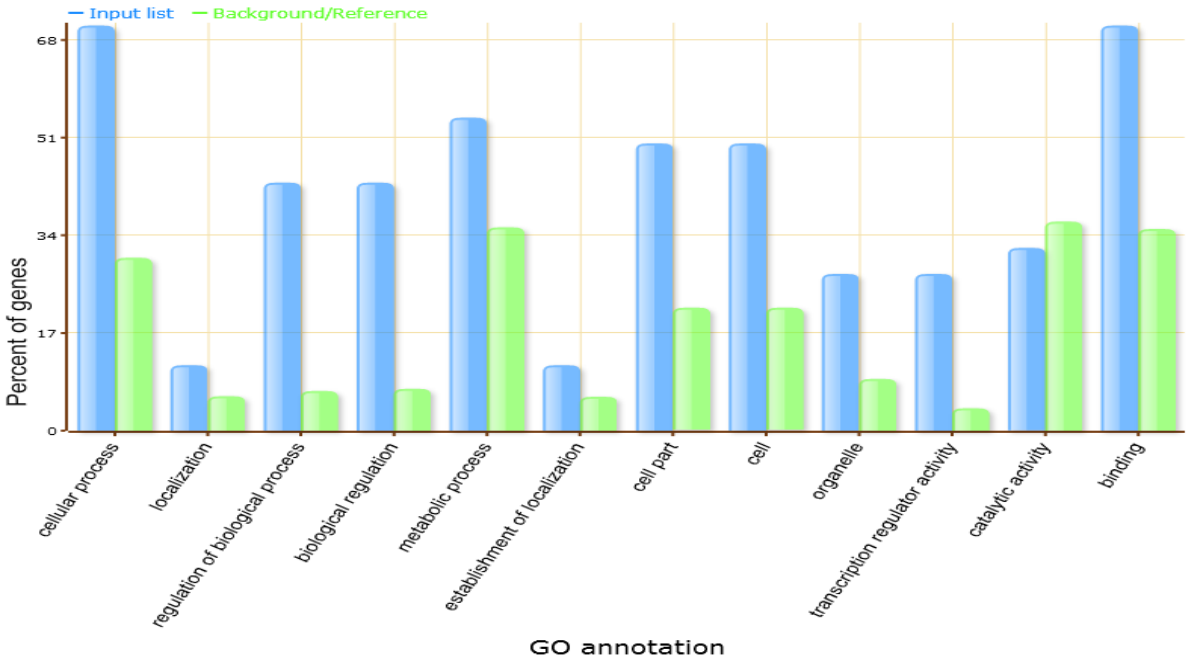

Figure 1. GO flash chart obtained from the agriGO analysis of the gene list

The genelist obtained from psRNATarget was provided as input to agriGO and the annotation was performed. The miRNAs were assigned to various cellular functions as given in Figure 1 and graphical representations of the biological processes were obtained using agriGO (Figure 2). The results showed 36 GO terms enriched for biological process, 19 for molecular function, and 10 terms for cellular component. In general, the GO terms that were significantly enriched are related to cellular and metabolic processes. The REVIGO analysis was carried out to visualize the GO results.

4. CONCLUSION

An attempt was made to develop a user friendly pipeline for analyzing NGS data for identifying miRNAs. The aim was to make use of publicly accessible platforms with graphical user interfaces which will have less computational complexity. The pipeline was developed by analyzing NGS data from *Brachypodium distachyon* under drought stress conditions. An analysis pipeline was developed using sRNAbench for read processing, miRNA sequence identification using miRBase, target prediction using psRNATarget and enrichment analysis using agriGO and REVIGO.

The differential expression analysis using sRNAbench generated a list of miRNAs that are differentially expressed which was used to find out the gene targets using psRNATarget. From the miRNA target analysis we obtained 118 target genes for the differentially expressed 32 miRNA sequences, from which a gene list was prepared. The results showed many GO terms enriched for biological processes, molecular functions and cellular components. The new pipeline has successfully identified miRNAs related to primary and secondary metabolic process, catabolic process, carbohydrates metabolic process and cellular homeostasis.

ACKNOWLEDGEMENTS

The authors thank the Director, JNTBGRI for providing necessary facilities to carry out the work.

REFERENCES

[1] Victor M., MicroRNAs in differentiating tissues of Populus and Eucalyptus trees, 2006, Dissertation - University of Pretoria. http://repository.up.ac.za/handle/2263/26140

[2] Lau NC, Lim LP, Weinstein EG, Bartel DP., An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans,* Science, 2001; 294(5543):858-62.

[3] Adai A., Johnson C, Mlotshwa S, Archer-Evans S, Manocha V, Vance V, et al.,. Computational prediction of miRNAs in *Arabidopsis thaliana*, 2005; Genome research 15:78-91

[4] Zhao Y, Srivastava D.,A developmental view of microRNA function, Trends in Biochemical Sciences, 2007; 32:189-197.

[5] Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ., miRBase: tools for microRNA genomics, Nucleic Acids Research, 2008; 36(Database issue): D154-8.

[6] Hackenberg M, Sturm M, Langenberger D, Falcón-Pérez JM, Aransay AM., miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments, Nucleic Acids Research, 2009; 37(Web Server issue):W68-76. doi: 10.1093/nar/gkp347

[7] Hackenberg M, Rodríguez-Ezpeleta N, Aransay AM., miRanalyzer: an update on the detection and analysis of microRNAs in high-throughput sequencing experiments, Nucleic Acids Research, 2011; 39(Web Server issue):W132-8. doi: 10.1093/nar/gkr247

[8] Qian K, Auvinen E, Greco D, Auvinen P., miRSeqNovel: an R based workflow for analyzing miRNA sequencing data, Mol Cell Probes, 2012; 26(5):208-11. doi: 10.1016/j.mcp.2012.05.002.

[9] Bertolini E, Verelst W, Horner DS, Gianfranceschi L, Piccolo V, Inzé D, et al., Addressing the Role of microRNAs in Reprogramming Leaf Growth during Drought Stress in *Brachypodium distachyon,* Molecular Plant, 2013; 6(2): 423–443. doi:10.1093/mp/sss160