An exhaustive search for intronic microRNAs in the cassava genome

Amika Yawichai[1], Tanawut Srisuk[2], Saowalak Kalapanulak[2,3], Chinae Thammarongtham[4*], Treenut Saithong[2,3*]

[1]*Bioinformatics and Systems Biology Program, School of Bioresources and Technology, and School of Information Technology, King Mongkut's University of Technology Thonburi (Bang Khun Thian), Bangkok, 10150, Thailand*
[2]*Systems Biology and Bioinformatics Research Group, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi (Bang Khun Thian), 10150, Thailand*
[3]*School of Bioresources and Technology, King Mongkut's University of Technology Thonburi (Bang Khun Thian), Bangkok, 10150, Thailand*
[4]*Biochemical Engineering and Pilot Plant Research and Development Unit, National Center for Genetic Engineering and Biotechnology at King Mongkut's University of Technology Thonburi (Bang Khun Thian), 10150, Thailand*
**e-mail: Chinae@biotec.or.th; Treenut.sai@kmutt.ac.th*

Abstract

Intronic mircroRNAs are a class of miRNAs which locate in intronic regions of any genes. They are short RNAs which are about 18-25 nucleotide long. Similar to the other class of miRNAs, intronic miRNAs have been known as a gene regulator relevant to cell growth and development, and cell response to stresses. According to their significant roles in cellular regulation, experimental as well as computational studies have been performed by aiming at extensive identification of the miRNA in species of interest. However, the finding of intronic miRNA is still limited. In this work, we thus used an *ab initio* approach to exhaustively search for intronic miRNAs in cassava, a starchy root crop. The study identified 127,485 putative intronic miRNAs partially overlapped with the result of deep sequencing method from the independent study. With the target analysis of these intronic miRNAs, it was suggested that the putative intronic miRNAs involved in essential cellular activity maintenance and stress response.

Keywords: intronic miRNA, cassava, computational identification

Introduction

Decades of research on non-coding RNAs (ncRNAs) has turned the meaning of the non-coding region in a DNA sequence, from junk to a valued message inherited from the ancestors. MicroRNA (miRNA) is a class of ncRNAs that has been studied at most during the last decade. In plant, miRNAs are mainly located in intergenic regions (Tong et al. 2013). Most of them were firstly transcribed to very long primary miRNAs (pri-miRNAs) by RNA polymerase II enzyme. The transcripts that form stem-loops were then cleaved by Dicer-like1 enzyme (DCL1) in plants, and the remaining stem-loop precursor miRNAs (pre-miRNAs) were subsequently cleaved to 18-25 nucleotide (nt) long miRNA/miRNA* duplexs (Monteys et al. 2010). The miRNAs are well known to regulate the fate of mRNA towards the translation through degradation process, whereby the guide strand of the duplex miRNAs complementary bound to the targeted mRNA resulting in the degradation or target translation inhibition (Saikumar and Kumar 2014). For the intronic miRNAs, some of their transcription related to the splicing mechanism of mRNAs (Tong et al. 2013).

Many novel miRNAs have been identified in broad organisms where they were reported to play an important role in diverse cellular regulations, including cell growth, development, and differentiation (Liu et al. 2012; Dong et al. 2013). One interesting example of miRNA function

is plants was for miR397 (Zhang et al. 2013). Overexpression of miR397 was reported to promote grain size and panicle branching in rice. In addition, miRNAs have been evidently involved in the cellular responses to biotic and abiotic stresses. miR939 was found to silence auxin signaling when host cells were infected, resulting pathogen starvation (Grant and Joes 2009). Another example is miR159, which involves in seed germination under abiotic stresses through regulating abcisic acid (ABA) phytohormone (Saikumar and Kumar 2014).

Conventional approaches have successfully identified miRNAs over a long period. However, advanced technology-aid approaches have been introduced to overcome the limitations of the current methods. With the advents of genome technology together with the thriving analytical techniques, many computational tools were developed for miRNA identification (Gomes et al. 2013). These tools were classified into two groups: (i) homology-based approaches and (ii) *ab initio* approaches. (Allmer and Yousef 2012). Homology-based approaches take the advantage of the available known miRNAs data, which are used for inferring the miRNA from the genome of interest. Without the background knowledge, these approaches are unable to perform a prediction (Allmer and Yousef 2012; Tempel and Tahi 2012; Gomes et al. 2013). On the contrary, *ab initio* approaches predicte the miRNAs from the characteristic features of miRNAs. Independent of the known miRNA datasets. These methods, thus, are able to provide the prediction of a novel miRNA through a machine learning protocol (Allmer and Yousef 2012; Lertampaiporn et al. 2013).

"HeteroMirPred" is a recently published computational tool that helps predict miRNA based on *ab initio*. It constructed miRNA models from heterogeneous ensemble method, and later exploited them in the miRNA prediction. An overt advantage of HeteroMirPred tool over the other *ab initio* -based tools is, the uses of multiple algorithms to prevent the false positive (FP) prediction resulting from the over-fitting of single algorithms. Furthermore, the features included in this tool were claimed to sufficiently distinguish miRNAs from other RNA species (Lertampaiporn et al. 2013).

Cassava (*Manihot esculenta* Crantz) is a staple crop plant, whose significance for being a carbohydrate source is at the world-class level. The huge demand on cassava starch is not only for food and feeds, but also for raw materials in a wide-range of industries: fuel, paper, textile, and adhesive (Howeler et al. 2013). Accordingly, the main focus on cassava research is to answer the question of sufficient production to serve the growing demand in the future. miRNA which play an important role in plant stress responses, has been considered as one key factors for crop yield improvement (Zhou and Luo 2013). However, the available knowledge of miRNAs in plants including cassava is very few, relative to their significance in controlling the yield of such crops under both biotic and abiotic stresses.

Availability of cassava genome sequence supporting by the effective bioinformatics techniques has made the computational prediction of cassava miRNAs possible. In this study, miRNAs of cassava were predicted from the exiting genome sequence (Prochnik et al. 2012), focusing only on the intronic miRNA. HeteroMirPred tool was selected to screen putative miRNAs of cassava from intronic sequences. The predicted results yielded 132,732 pri-miRNAs which were subsequently compared with the set of cassava intronic miRNAs identified by deep sequencing (Ballén-Taborda et al. 2013). Interestingly, the identified cassava miRNA herein may involve growth and development of cassava.

Methodology

Data sets

Plant genomes. Cassava (*Manihot esculenta* Crantz) genome, transcript and annotation information (version 4.1) were downloaded from Phytozome database version 9.1

(http://www.phytozome.net/), and *Arabidopsis thaliana* annotation information (release 10) was downloaded from Phytozome database version 10.
Published plant miRNA data. The published plant miRNA sequences were collected from two databases; 5,475 mature miRNAs belonging to 74 species in kingdom of Virisiplantae from miRBase Release 20 (http://www.mirbase.org/), and 597 mature miRNAs belonging to 128 species from PMRD (plant microRNA database;- http://bioinformatics.cau.edu.cn/PMRD/).

Softwares

HeteroMirPred (Lertampaiporn et al., 2013), the pre-miRNA predictor was used for identifying the putative miRNAs in the genome. The predicted miRNAs were folded using Mfold version 3.6 (Zuker et al. 2003) to evaluate the appropriateness of the pre-miRNA secondary structure. Blast (Nucleotide-Nucleotide BLAST 2.2.28+) was employed to assess the conservation of the predicted putative miRNAs. psRNATarget (http://plantgrn.noble.org/psRNATarget/) and Targetfinder 1.6 (Fahlgren et al. 2007) were used in the miRNA target prediction, in which the target genes of the miRNAs were classified based on biological network gene ontology (Bingo) application on Cytoscape version 3.1.1.

Prediction of intronic-miRNAs in cassava

Sequences of intron regions extracted from cassava genome were fragmented into 120-nucleotide sliding windows with 20-nucleotide overlapping (Lertampaiporn et al. 2013; Xuan et al. 2011). The stem-loop, inferred from the secondary structures folded by Mfold, was identified according to the criteria presented by Thakur et al., (2011). The stem-loop containing sequences were predicted as a putative pre-miRNA, which were then undergone the conservation analysis by blasting against the plant mature miRNAs using the criteria of Meyers et al., (2008).

Prediction of miRNA targets and Gene ontology (GO) enrichment analysis

Potential targets of the putative miRNAs were predicted through the consensus results from two tools: psRNATarget and Targetfinder 1.6. The following parameters were used in prediction:

1) For psRNATarget; maximum expectation = 3.0, hpsize = 20, target accessibility score = 25.0, flanking length around target site = 17 bp for upstream and 13 bp for downstream, and range of central mismatch ~ 9-11 nucleotides.
2) For Targetfinder; prediction score cutoff value = 4.

All predicted target genes of the putative intronic miRNAs were categorized based on functional enrichment using Bingo.

Results and Discussion

Putative intronic miRNAs in cassava genome sequence

In this study, we exploited a computation pipeline (**Figure 1**) to identify intronic miRNAs from cassava genome sequence. A total of 716,588 intronic regions of cassava genome were screened for miRNAs by HeteroMirPred. After merging the overlapped regions, 132,732 sequences which contain pre-miRNAs were acquired. Many miRNAs are evolutionary conserved in among plant species (Zhang et al. 2006; Jones-Rhoades 2012), consequently, homology based approach were employed to classify for conserved miRNAs (Griffiths-Jones

2004). Since mature miRNAs were processed from one arm of pre-miRNAs, therefore stem-loop structures were a one of the important features for mature miRNA identification (Thakur et al. 2011, Meyers et al. 2008). pri-miRNAs were folded to stem-loop structures to select pre-miRNAs based on pre-miRNA structure characteristics described by Thakur et al., (2011). Based on secondary structure prediction, 127,485 sequences containing stem-loop structures were obtained from 112,770 putative pre-miRNAs. These stem-loop containing sequences were then searched against published plant mature miRNAs by blast to categorize them into two groups: conserved and non-conserved miRNAs. There was only one putative pre-miRNA that matched with a known miRNA sequence, miR398. A stress-responsive miRNA, miR398, was suggested to be involved in biotic and abiotic stress response in plants (Zhu et al. 2011). All other putative pre-miRNAs which did not match were assigned as non-conserved miRNAs.

However, there were 18,318 miRNA predicted sequences that their stem-loop structures could not be obtained. Generally, miRNAs needed to be folded into stem-loop structures before continuing process and functions (Saikumar and Kumar 2014). Many available miRNA prediction tools used this characteristic feature to distinguish miRNA from other non-coding RNAs (ncRNAs) by considering from their secondary structures (Washietl et al. 2012; Rivas and Eddy 2000; Hamilton and Davis 2007). For HeteroMirPred, it is not only secondary structure feature, which is used in identifying pre-miRNAs but also other selected features, particularly "the Self Containment (SC)-derived features. Such features represent the intrinsic robustness of real stem-loop structures of pre-miRNAs (Lertampaiporn et al. 2013). Consequently, some predicted sequences from HeteroMirPred cannot fit to secondary structure criteria.

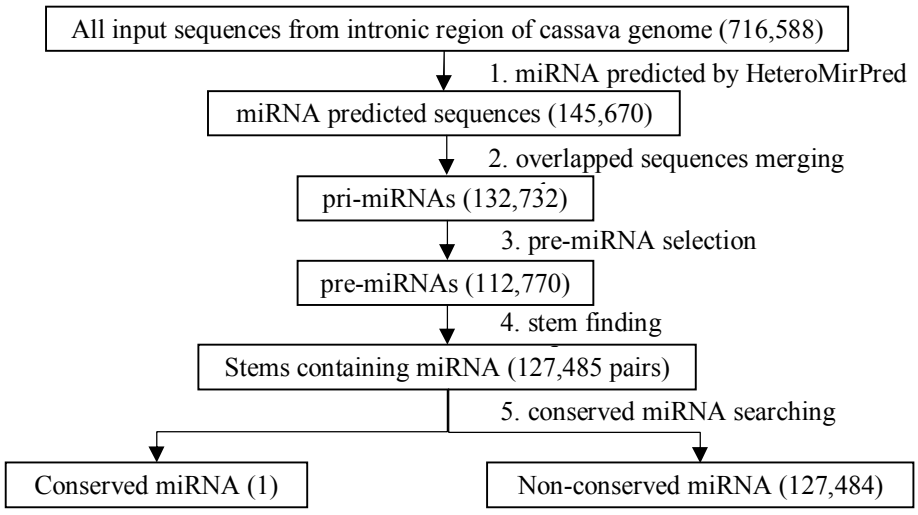


Figure 1: Intronic miRNAs identification work flow. The numbers in parenthesis are obtained miRNAs in each step.

To support our prediction, we compared the resulting putative miRNAs with the deep sequencing data of cassava (Ballén-Taborda et al. 2013). It showed that 26 of obtained intronic miRNAs including one conserved intronic miRNA were matched. The statistical analysis was applied to indicate the level of confidence of the prediction result. To demonstrate the probability for obtaining 26 or more miRNAs from randomly selecting 131,080 cassava intronic miRNAs without replacement from the total of 716,588 all input sequences, hypergeometric probability for obtaining equal or greater miRNA were calculated following the formula below.

$$p(x \geq 26) = 1 - p(x \leq 25) \quad \textbf{(1)}$$

$$p(x \leq 25) = \sum_{i=0}^{25} \frac{\binom{60}{25-i}\binom{716{,}528}{131{,}055+i}}{\binom{716{,}588}{131{,}080}} = 0.999998181911 \quad \textbf{(2)}$$

Whereas, the number of "successes" in the population (k) is 60 intronic miRNAs from deep sequencing, the number of "successes" in the sample (x) is 26 intronic miRNAs that were corresponded with intronic miRNAs from deep sequencing, the size of the population (N) is 716,588 input sequences from intronic regions, and the number sampled (n) is 131,080 intronic miRNAs from this study. The hypergeometric probability showed that the 26 intronic miRNAs we predicted from all input sequences is a rare event, indicating the power of the prediction method.

For the miR398 homologue predicted from our pipeline, we found mature miRNAs on both arm of pre-miRNA while deep sequencing can identify only one mature miRNA (**Figure 2**). In fact, one pre-miRNA can encode more than one mature miRNA (Tong et al. 2013; Griffiths-Jones et al. 2006). It is possible that both mature miRNAs independently function in different conditions.

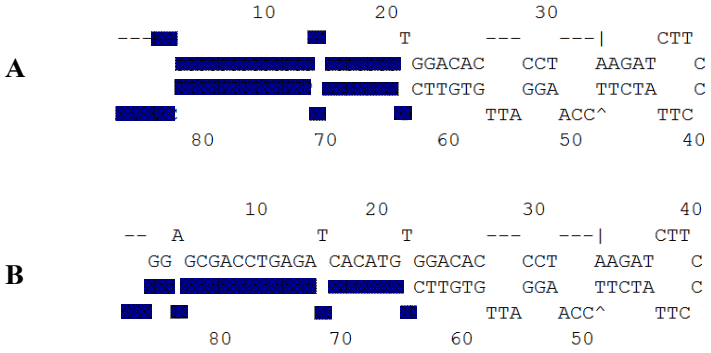


Figure 2: miR398 structures. The yellow highlights on pre-miRNAs structures are mature miRNA locations. The secondary structures of miR398 homologue predicted in this work (**A**) and miR398 identified from deep sequencing data (**B**) were showed.

Despite the positive results obtained, it need to be noted that computational approaches might include a huge FP in the results (Lertampaiporn et al. 2013). Therefore, experimental validation is required to confirm these results.

Gene targets and function analysis

miRNAs regulate gene expression by complementarily binding against targets gene. The biogenesis of miRNAs revealed that mature miRNA, a short sequence (~18-25 nt) from one arm of pre-miRNA formed complex with RISC (RNA-Induced Silencing Complex) and degraded or repressed gene target by degree of perfectly binding between mature miRNA and their target (Dai et al. 2010). Therefore, function of each miRNA can be investigated by gene that it bound. However non-conserved miRNA has no any mature miRNA report, therefore in this work gene targets were predicted from stems of pre-miRNAs.

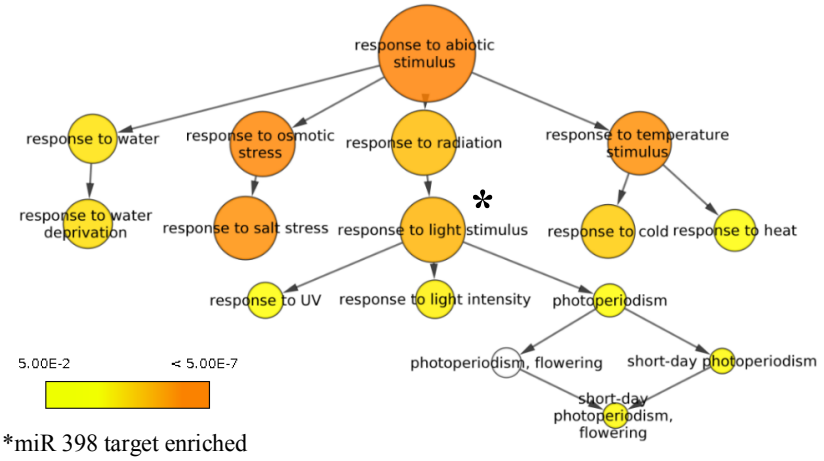
Almost available target predictors always set the cut off value that specific for plant model, *A. thaliana,* which may not be suitable for non-model plants such as cassava (Srivastava et al. 2014). Srivastava et al., (2014) found that the combination of psRNAtarget and Targetfinder can improve performance by increasing true-positive (TP) for non-plant model organisms.

Therefore we employed these two tools for gene targets prediction. 4,272 gene targets from 1,189 intronic miRNAs were obtained from the consensus result the two tools.

GO analysis revealed that mostly gene target were significant enriched in ontology term of gene target associated with maintenance of basic cellular function *e.g.* transcription factors (**Table 1**). In addition, there are many reports on miRNAs contributed plant survival under adverse environmental conditions (Contreras-Cubas et al. 2012; Saikumar and Kumar 2014; Xia et al. 2014). miR398 is a miRNA proposed to be directly linked to the plant stress regulatory network and regulates plant responses to oxidative stress, water deficit, salt stress, abscisic acid stress, ultraviolet stress, copper and phosphate deficiency, high sucrose and bacterial infection. This review highlights recent progress in understanding the crucial role of miR398 in plant stress responses, and also includes a discussion of miR398-centered gene regulatory network. Interestingly, a gene target of miR398 in this work, phytochrome interacting factor 3 (PIF3) founded to involve in response to light stimulus (**Figure 3**). This observation was corresponding to a role of miR318 in *Arabidopsis* (Park et al. 2004; Liu et al. 2013), PIF3 was showed negatively control chlorophyll biosynthesis and photosynthesis in etiolated seedling. Therefore, it is possible that the presence of miR398 may increase the ability of cassava growth in the darkness.

Table 1 Gene ontology (GO) enrichment analysis: Top 10 ranks of significant target genes enriched. M=Molecular function, C=Cellular component, and B=Biological process.

GO term	p-value	Gene number	Total gene number in GO term	Description	Ontology
44464	3.98E-87	462	11708	cell part	C
44238	2.34E-67	534	5719	primary metabolic process	B
44237	3.40E-60	501	5407	cellular metabolic process	B
166	5.56E-56	337	2085	nucleotide binding	M
16740	4.69E-41	338	2429	transferase activity	M
6793	1.83E-40	157	980	phosphorus metabolic process	B
1882	4.62E-37	227	1393	nucleoside binding	M
43170	4.82E-36	366	4086	macromolecule metabolic process	B
43227	2.82E-23	243	5767	membrane-bounded organelle	C
16787	9.93E-23	305	2632	hydrolase activity	M



*miR 398 target enriched

Figure 3: GO enrichment analysis of target genes which were responding to stimulus. The colors of circle represented gene significant in GO-terms and the circle sizes represented gene number which enriched in GO-terms.

However, the identification of non-conserved miRNAs by this method may lose many possible gene targets because: 1) these tools can only detect the end part of the sequence and 2) the selection of consensus gene targets may lost some TP results from each tools.

Conclusions

miRNA identification is an active research in range of organisms, including plants. The studies in plants are mostly based on *Arabidopsis* and other model plants such as rice. Only a recent publication presented the study in cassava, though it is the obviously significant crops era. In this study, we used HeteroMirPred, an *ab initio* method, to exhaustively screen for miRNA in cassava, focusing on intronic miRNA-class of miRNA. Our study provided set of 127,485 putative intronic miRNAs, one of which, miR398 homologous, was predicted to possess interesting function in normal growth of cassava. To increase the contribution of this finding, experimental validation, probably using high throughput technology, should be conducted.

Acknowledgements

The authors would like to extent our deep gratitude to National Center for Genetic Engineering and Biotechnology (BIOTEC) for Amika Yawichai's scholarship.

References

Allmer J., and Yousef M. (2012). Computational methods for *ab initio* detection of microRNAs. *Frontier in Genetics,* 3, 1-5.

Ballén-Taborda C., Plata G., Ayling, S., Rodríguez-Zapata, F., Lopez-Lavalle L.A.B., L.A., Duitama, J., and Tohme, J. (2013). Identification of Cassava MicroRNAs under Abiotic Stress. *International Journal of Genomic*, 1-10.

Contreras-Cubas C., Palomar M., Arteaga-Vázquez M., Reyes J.L., and Covarrubias A.A. (2012). Non-coding RNAs in the plant response to abiotic stress. *Planta,* 236, 943-958.

Cuperus J.T., Fahlgren N., and Carrington J.C. (2011). Evolution and Functional Diversification of MIRNA Genes. *The Plant Cell*, 23, 431–442.

Dai X., and Zhao, P.X. (2011). psRNATarget: A Plant Small RNA Target Analysis Server. *Nucleic Acids Research*, 1-5.

Dong H., Lei J., Ding L., Wen Y., Ju H., and Zhang X. (2013). MicroRNA: Function, Detection, and Bioanalysis. *Chemical Review,* 6207-6232.

Fahlgren N, Howell M.D., Kasschau K.D., Chapman E.J., Sullivan C.M., Cumbie J.S., Givan S.A., Law T.F., Grant S.R., Dangl J.L., Carrington J.C. (2007). High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. *PLoS ONE*, 2.

Gomes C.P.C., Cho J.-H., Hood L., Franco O.L., Pereira R.W., and Wang K. (2013). A Review of Computational Tools in microRNA Discovery. *Frontiers in genetics*, *4*(May), 81.

Griffiths-Jones S., (2004). The microRNA Registry. *Nucleic Acids Research,* 32, 109 111.

Griffiths-Jones S., Grocock R.J., Dongen S., Bateman A., and Enright A.J. (2006). miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Research,* 34.

Hamiton R.S., and Davis I. (2007). RNA localization signals: Deciphering the message with bioinformatics. *Seminars in Cell & Developmental Biology*, 18, 178-185.

Howeler R., Lutaladio N., and Thomas G. (2013). Save and Grow: Cassava. *Food and agriculture organization of the united nations*, Rome.

Jones-Rhoades M.W. (2012). Conservation and divergence in plant microRNAs. *Plant Molecular Biology,* 80(1), 3-16.

Lertampaiporn S., Thammarongtham C., Nukoolkit C., Kaewkamnerdpong B., and Ruengjitchatchawalya M. (2013). Heterogeneous ensemble approach with discriminative features and modified-SMOTEbagging for pre-miRNA classification. *Nucleic acids research*, *41*(1), e21.

Liu J, Githinji J, Mclaughlin B, Wilczek K, and Nolta J. (2012). Role of miRNAs in neuronal differentiation from human embryonic stem cell-derived neural stem cells. *Stem Cell Reviews and Reports,* 8(4), 1129-1137.

Meyers B.C., Axtell M.J., Bartel B., Bartel D.P., Baulcombe D., Bowman J.L., Cao X., Carrington J.C., Chen X., Grenn P.J., Griffiths-Jones S., Jacobsen S.E., Mallory A.C., Martienssen R.A., Poethig R.S., Qi Y.,Vaucheret H., Voinnet O., Watanabe Y., Weigel D., and Zhu J. (2008). Criteria for Annotation of Plant MicroRNAs. *The Plant Cell,* 20, 3186-3190.

Monteys A.M., Spengler R.M., Wan J., Tecedor L., Lennox K.A., Xing Y., and Davidson B.L. (2010). Structure and activity of putative intronic miRNA promoters. *RNA,* 16, 495-505.

Patanim O., Lertpanyasampatha M., Sojikul P., Viboonjun U. and Narangajavana J. (2012). Computational Identification of MicroRNAs and Their Targets in Cassava (*Manihot esculenta* Crantz). *Molecular Biotechnology.*

Prochnik S., Marri P. R., Desany B., Rabinowicz P. D., Kodira C., Mohiuddin M., and Rounsley S. (2012). The Cassava Genome: Current Progress, Future Directions. *Tropical plant biology*, *5*(1), 88–94.

Rivas E., and Eddy S.R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioiformatics,* 16(7), 583-605.

Saikumar K., and Kumar V.D. (2014). Plant MicroRNAs: An Overview. *Agricultural Bioinformatics,* 139-159.

Srivastava P.K., Moturu T.R., Pandey P., Baldwin I.T., and Pandey S.P. (2014). A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction. *BMC Genomics.* 15:348, 1471-2164.

Temple S., and Tahi F. (2012). A fast *ab-initio* method for predicting miRNA precursors in genomes. *Nucleic Acids Research,* 40(11).

Thakur V., Wanchana S., Xu M., Bruskiewich R., Quick W.P., Mosig A., and Zhu X. (2011). Characterization of statistical features for plant microRNA prediction *BMC Genomics*, 12(108), 1471-2164.

Tong Y., Peng H., Zhan C., Fan L., Ai T., and Wang S. (2013). Genome-Wide Analysis Reveals Diversity of Rice Intronic miRNAs in Sequence Structure, Biogenesis and Function. *PLOS ONE,* 8(5), 1-12.

Tonukari N.J. (2004). Cassava and the future of starch. *Journal of Biotechnology,* 7(1).

Washietl S., Will1 S., Hendrix D.A., Goffl L.A., Rinn J.L., Berger B., and Kellis M. (2012). Computational analysis of noncoding RNAs. *WIREs RNA.*

Xuan P., Guo M., Liu X., Huang Y., Li W., and Huang Y. (2011). PlantMiRNAPred: efficient classification of real and pseudo plant pre-miRNAs. *Bioinformatics (Oxford, England)*, *27*(10), 1368–1376.

Zhang B., Pan X., Cannon C.H., Cobb G.P. and Anderson T.A. (2006). Conservation and divergence of plant microRNA genes. *The Plant Journal,* 46, 243-259.

Zhang Y., Yu Y., Wang C., Li Z., Liu Q., Xu J., Liao J., Wang X., Qu L., Chen F., Xin P., Yan C., Chu J., Li H., and Chen Y. (2013). Overexpression of microRNA OsmiR397 improves rice yield by increasing grain size and promoting panicle branching. *Nature Biotechnology,* 31, 848–852.

Zhou M., and Luo H. (2013). MicroRNA-mediated gene regulation: potential applications for plant genetic engineering. *Plant Molecular Biology,* 83(1-2), 59-75.

Zhu C., Ding Y., Liu H. (2011). MiR398 and plant stress responses. *Physiol. Plant.* 143:1-9.