Genome-wide investigation of lincRNAs in cassava and their potential roles in specific tissue

Rungaroon Suksamran[1*], Treenut Saithong[2, 3],
Chinae Thammarongtham[4], and Saowalak Kalapanulak[2, 3]

[1]Biotechnology Program, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi (Bang Khun Thian), Bangkok, Thailand
[2]Bioinformatics and Systems Biology Program, School of Bioresources and Technology, King Mongkut's University of Technology Thonburi (Bang Khun Thian), Bangkok, Thailand
[3]Systems Biology and Bioinformatics Research Group, Center for Agricultural Systems Biology, Pilot Plant Development and Training Institute, King Mongkut's University of Technology Thonburi (Bang Khun Thian), Thailand
[4]Biochemical Engineering and Systems Biology Research group, National Center for Genetic Engineering and Biotechnology at King Mongkut's University of Technology Thonburi (Bang Khun Thian), Thailand
[*]Corresponding author, E-mail: rungaroon.poy@gmail.com

Abstract

Long non-coding RNAs (lncRNAs) are an exclusive type of ribo-regulators controlling post-transcriptional regulation in various organisms, especially in plants. Recently, important functions of long intergenic non-coding RNAs (lincRNAs), non-coding transcripts with size > 200 nts and transcribed from the intergenic regions of the genome, such as flowering, ripening, fertility, and storage organ developments were reported in plants. However, the knowledge of cassava lincRNAs in particular tissues are still limited. In this work, we aim to do genome-wide identification of lincRNAs in 11 cassava tissues, including shoot, root, and callus by using tissue-specific RNA-seq data and bioinformatics approach. Here, we showed 1,258 lincRNAs with various expression levels in cassava tissues. Of these, 67 lincRNAs were commonly expressed in all 11 cassava tissues and 4 of them were proposed as potential reference genes for relative gene expression quantification measurement in cassava. Moreover, 657 lincRNAs specifically expressed in single individual tissues. *Cis*-regulatory target prediction suggested the potential roles of two lincRNAs involving in carbohydrate metabolism in storage roots and photosynthesis in cassava leaves. The genome-wide collection of lincRNAs in 11 specific tissues of cassava will be useful for further molecular function validation.

Keywords: *Cassava, Genome-wide, LincRNAs, Tissue-specific, Cis-regulatory target, Carbohydrate metabolism, Photosynthesis*

1. Introduction

Long non-coding RNAs (lncRNAs) are transcripts longer than 200 nucleotides (nts) without protein-coding capacity. Even their low expression level and low primary sequence conservation, when compared to protein-coding genes, their functions involving in several cellular processes have made them become the meaningful sequences in the genomes. LncRNAs can originate from the spliced intron of mRNAs (intronic-lncRNAs), antisense transcripts of coding genes (lncNATs) or independent transcription at intergenic regions (lincRNAs) and play important roles in post-transcriptional regulation of protein-coding genes (Yao, Wang, & Chen, 2019).

Long intergenic non-coding RNAs (lincRNAs) have been reported as a majority of plant lncRNAs (Zhao et al., 2019; Wang et al., 2019a) and conferred gene regulation in both *cis*- (regulating target genes located in the same genomic locus) and *trans*- (regulating target genes located in the different genomic locus) regulation. Overexpression of lncRNA23468 caused the increase of disease-resistant gene (NBS-LRRs) expression through miR482b target mimic and promoted disease resistance in tomato after fungal infection (Jiang et al., 2019). Wang et al., (2014) proposed a lincRNA named HID1 involving in the photomorphogenesis (light-mediated seedling development) through the protein-RNA complex in Arabidopsis. The silencing of HID1 promoted the increase of photomorphogenesis repressor, PIF3, activity and resulting in longer hypocotyls (Wang et al., 2014). These suggested the obvious evidence of lincRNA's role in plant disease resistance and development through *trans*-regulation.

For *cis*-regulation, Wang et al., (2019b) proposed the function of lincRNAs involving in heat tolerance in Chinese cabbage by using transcriptome analysis. They found differentially expressed lincRNAs together with their nearby protein-coding genes encoding for heat stress transcription factors (HSFs) (Wang et al., 2019b). Cold-related lincRNAs were identified in grapevine. Some of these lincRNAs showed differential expression responding to cold stress condition and their *cis*-regulatory target (nearby) genes were reported as plant stress response genes. For example, the expression level of lincRNA XLOC_019156 was significantly increased while its *cis*-regulatory target gene encoding for peroxisome biogenesis protein (VIT_202s0154g00610) was down-regulated under cold stress condition (Wang et al., 2019a). These pieces of evidence increase the attention of researchers for exhaustive lincRNA exploration in other plants (Hao et al., 2015; Golicz, Singh, & Bhalla, 2018).

Recently, several tissue-specific lincRNAs were investigated in various plant species. These observations suggested the potential roles of lincRNA in specific functions on plant developmental processes. For instance, a high correlation of expression level between lincRNAs and their corresponding mRNA targets involving in plant flowering was found in chickpea suggesting the function of lincRNA involving in flower developmental process (Khemka et al., 2016). Knockout of a lincRNA, XLOC_057324, resulting in decreasing in rice fertility. This suggested the involvement of lincRNA in rice sexual reproduction (Zhang et al., 2014). Varshney and colleagues proposed a potential role of lincRNAs involved in aroma biosynthesis of black tea since these lincRNAs showed differential expression in aroma specific tissue and shared the miRNA binding region with mRNA (target mimicry) of protein-coding gene involving in the biosynthesis pathway of volatile compound (Varshney et al., 2019). Wang et al., (2018) reported the involvement of lincRNA in fruit ripening as observing the co-transcriptional expression of lincRNA (XLOC_055641) with an adjacent protein-coding gene (Solyc06g051800) which is a target of ethylene regulation and specifically expressed in ripening fruit. Moreover, both of them were up-regulated during the ripe stage in tomato (Wang et al., 2018). The knowledge of functional lincRNA in particular plant tissues will encourage more understanding of plant post-transcriptional gene regulation and be an alternative resource beyond protein-coding genes for the next generation of functional genomics studies.

Nowadays, a collection of lincRNAs expressed in specific tissues of the most important cash crops for Thailand, cassava, is still limited, contrasting to its economic value. Only a few genome-wide transcriptome studies of protein-coding genes in cassava under tissue-specific conditions were performed (Wilson et al., 2017). Moreover, the investigation of lincRNAs in specific tissues of cassava is relatively less. A collection of lincRNAs in a specific tissue of cassava is not only benefited for cassava yield capacity improvement but also other molecular applications. For example, lincRNAs that associate with cassava storage root development or photosynthesis in leaf might shed some light on improving cassava storage root yield or enhance photosynthesis efficiency, respectively. Therefore, we put an effort to do a genome-wide screening of lincRNAs under different tissues of cassava plant using a bioinformatics approach and RNA-seq data from Wilson et al (2017). This provided a collection of lincRNAs in 11 cassava tissues including the shoot and root part of the cassava plant, and callus of cassava tissue culture. We investigated both common lincRNAs expressed in all 11 cassava tissues and tissue-specific lincRNAs. This genome-wide collection of lincRNAs in each cassava tissue might be beneficial of further tissue-specific functional validation or applying for tissue-specific biomarkers.

2. Objectives

This work aims to do a genome-wide screening of lincRNA in 11 cassava tissues by using RNA-seq data integrating with a bioinformatics approach.

3. Materials and Methods

Cassava RNA-seq data and genome annotation

RNA-seq data of 11 cassava tissues (Wilson et al., 2017) including; shoot apical meristem (SAM), leaf, midvein, petiole, stem, lateral bud, storage root (SR), fibrous root (FR), root apical meristem (RAM), organized embryogenic structure (OES) and friable embryogenic callus (FEC), consisted of three biological

replicates for each condition (except SR, containing two biological replicates), were downloaded from NCBI database (BioProject No. PRJNA324539). This RNA-seq data were constructed based on poly-A tail RNA library, Illumina RNA sequencing platform with reads of 101 bps in length. Cassava reference genome, AM560v6.1 including protein-coding gene annotation was retrieved from the Phytozome12 database and used as a template for RNA-seq read mapping.

Bioinformatics approach for lincRNAs identification

Qualified reads with average Phred-quality score > 20 in each sample were mapped on the cassava reference genome using STAR aligner (Dobin et al., 2013). Aligned reads with uniquely mapped and identity > 70% were assembled using Cufflinks (Trapnell et al., 2010) and guided by cassava protein-coding gene annotation. Intergenic assembled transcripts were proceeded to identify lincRNAs by several filtering criteria as shown in figure 1 including (i) size > 200 nt, (ii) protein-coding gene distance > 500 nts (iii) no coding potential by CPC (Kong et al., 2007) and CPAT (Wang et al., 2013) and (iv) no similarity to protein sequences or domains. Expressed lincRNAs were identified from transcripts passed through the above-mentioned filtering criteria with mapped reads ≥ 10 in at least 2 replicated data sets.

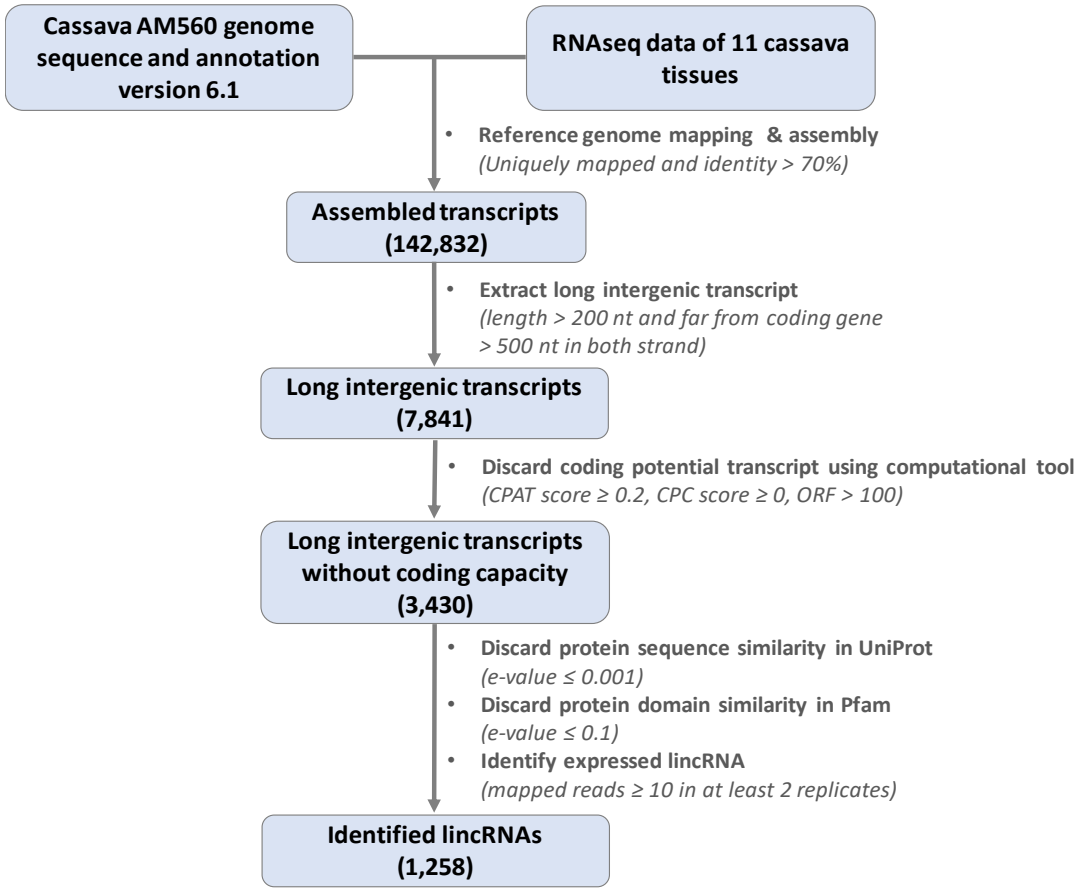

Figure 1 Workflow for lincRNA identification using a Bioinformatics approach

Identification of potential common and tissue-specific lincRNAs and cis-target prediction

Potential tissue-specific lincRNAs were identified from lincRNAs which expressed in only a specific tissue while common lincRNAs were found their expression in all 11 cassava tissues. Abundance estimation of identified lincRNA transcripts was quantified via reads counting with normalization by GeTMM (Smid et al., 2018). Target genes of interesting lincRNAs were predicted from protein-coding genes located within 10 Kb in up or downstream regions of lincRNAs.

4. Results and Discussion

Identified lincRNAs and their expression in particular cassava tissues

Over 454 million qualified reads from 11 tissue samples of cassava were mapped to the reference genome with more than 80% mapping rate for each sample. After transcript assembly via the bioinformatics pipeline, several numbers of lincRNAs were identified in each cassava tissues as shown in figure 2A. Finally, we found 1,258 unique lincRNAs expressed in at least one cassava tissue. Characteristics of identified lincRNAs based on length were presented as size distribution in figure 2B. The length of lincRNAs is in the range of 202 to 8,312 nt and the average was lower than that of 8,650 expressed protein-coding transcripts (102 to 16,325 nt in length). The identified lincRNAs showed lower exon numbers than protein-coding transcripts (figure 2C). In terms of expression level, lincRNAs showed significant differences and lower expression levels than protein-coding transcripts in all 11 tissues (Wilcoxon rank-sum test, p-value < 0.05) (figure 2D). Moreover, their expression levels of lincRNAs among 11 tissues were significantly different (Kruskal-Wallis test, p-value < 0.05). It suggested the diverse roles of lincRNAs in different tissues. The features of shorter in sizes, lower exon numbers and lower expression levels of identified lincRNAs comparing to those of protein-coding transcripts in cassava were consistent with the previous reports of lncRNA's characteristics in other plants (Golicz, Singh, & Bhalla, 2018; Wang et al., 2018).

Common lincRNAs in all cassava tissues and their potential roles as reference transcripts for relative expression quantification measurement

The 67 lincRNAs showed their expression in all 11 cassava tissues with variations. This suggests common lincRNAs that play various roles in all cassava tissues. Among of these common lincRNAs, four lincRNAs including TCONS_00133763, TCONS_00142409, TCONS_00097964, and TCONS_00077571 showed comparable expression levels with three current cassava reference genes (Hu et al., 2016): TUB, beta_6 tubulin (Manes.08G061700), 26Sproteasome regulatory complex subunit N10 (Manes.02G137500) and Nascent polypeptide-associated complex subunit beta (Manes.09G005100) as shown in figure 3. The mean expression level for four lincRNAs was in the range of 2.28 to 2.84, whereas the mean of the expression level of three current reference genes was in the range of 2.50 to 3.92. Low variation of their expression also observed for all four lincRNAs at the same level as current reference genes. This suggested the potential for using these four lincRNAs as reference transcripts for further cassava molecular studying such as the quantitative reverse transcription-polymerase chain reaction (RT-qPCR) method. To the best of our knowledge, there is no evidence of lncRNA utilization as a reference gene in the plant but recently reported as markers in cervical cancer (Iempridee et al., 2018).

Potential functional lincRNAs in specific leaves or storage roots tissue

A further investigation suggested that a total of 657 lincRNAs were expressed specifically in particular tissues as shown in figure 4A. This number (657) of tissue-specific lincRNAs in cassava is quite close to the previously reported number (641) in chickpea (Khemka et al., 2016). Comparing to all tissue-specific lincRNAs, almost 55% of them were found in FEC. When considering all expressed lincRNAs in each tissue, almost 46% of them are specific lincRNAs in FEC, while the other tissues showed a lower fraction of tissue-specific lincRNAs (figure 4B). This may suggest that the FEC tissue of cassava might require more specific control by lincRNAs than other tissues. This might be a reason why the numbers of expressed lincRNAs and fraction of tissue-specific lincRNAs in FEC are very high which is consistent with the previous reports of lncRNA's role in embryogenesis and cell differentiation (Fatica & Bozzoni, 2014; Huang et al.,

2019). A storage root- specific lincRNA, TCONS_00032023, and its *cis*- target, FASCICLIN- like arabinogalactan, was proposed in this work. This targeted gene was previously reported as its association with root growth and seed development in the plant (Basu et al., 2016). The expression levels of both were highest in cassava storage root samples (figure 4C). Besides, a leaf- specific lincRNA, TCONS_00069151 involving in the photosynthesis process through *cis*-regulated target gene encoding for complex photosystem II, was found to be expressed at the highest levels in leaf samples (figure 4D). These results suggest that our identified lincRNAs have their potential roles in cassava tissue-specific functions.

(A)

Type of tissue	Number of protein-coding transcripts	Number of lincRNAs
SAM	4,228 (4,218)	218
Leaf	4,639 (4,624)	251
Midvein	4,992 (4,978)	252
Petiole	5,074 (5,058)	279
Stem	5,027 (5,013)	275
Lateral bud	5,233 (5,219)	291
SR	3,879 (3,863)	276
FR	5,657 (5,640)	335
RAM	4,133 (4,124)	155
OES	5,697 (5,677)	499
FEC	5,302 (5,286)	785
Total	8,650 (8,623)	1,258

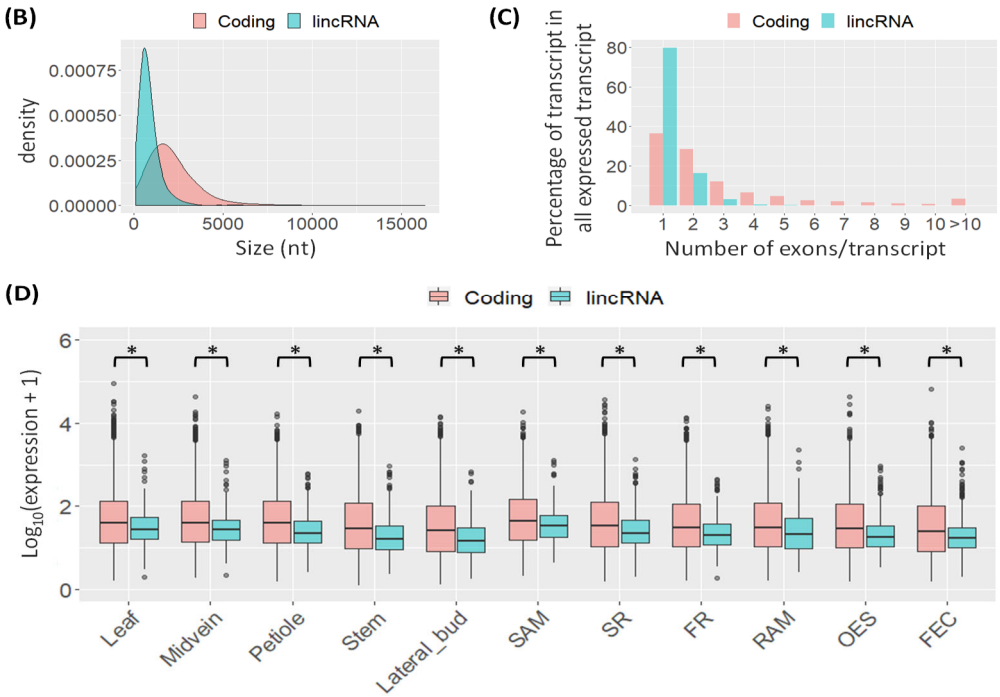

Figure 2 Identified lincRNAs and their characteristics. (A) The number of expressed lincRNAs and protein-coding transcripts in 11 cassava tissues. The number in the parenthesis indicates the number of expressed protein-coding genes. (B) The size distribution of identified lincRNAs comparing to protein-coding transcripts. (C) Distribution of exon numbers in identified lincRNAs comparing to protein-coding transcripts. (D) The expression level of lincRNAs comparing to protein-coding transcripts in 11 cassava tissues. *Denotes the statistically significant difference at $p < 0.05$, based on Wilcoxon rank-sum test

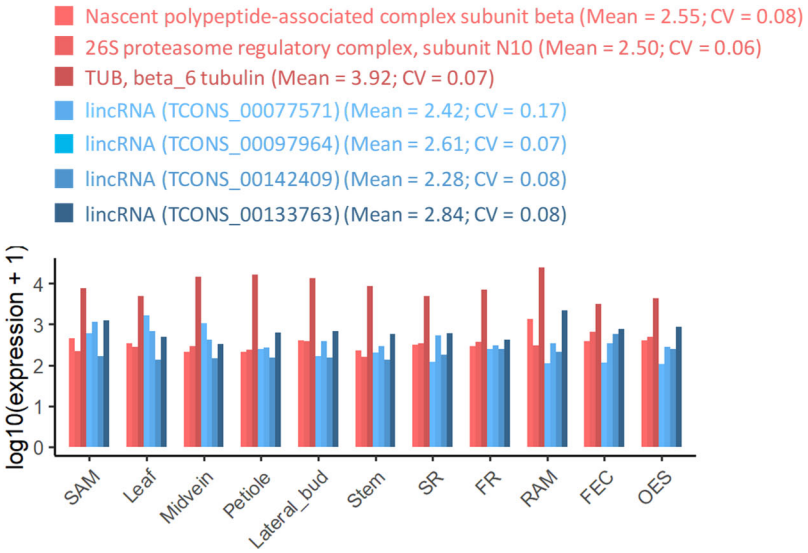

Figure 3 The four common lincRNAs (blue labels), proposed as reference transcripts for relative expression quantification measurement and three current reference genes (red labels) with their expression level across 11 cassava tissues. Mean is the average expression level and CV is the coefficient of variation

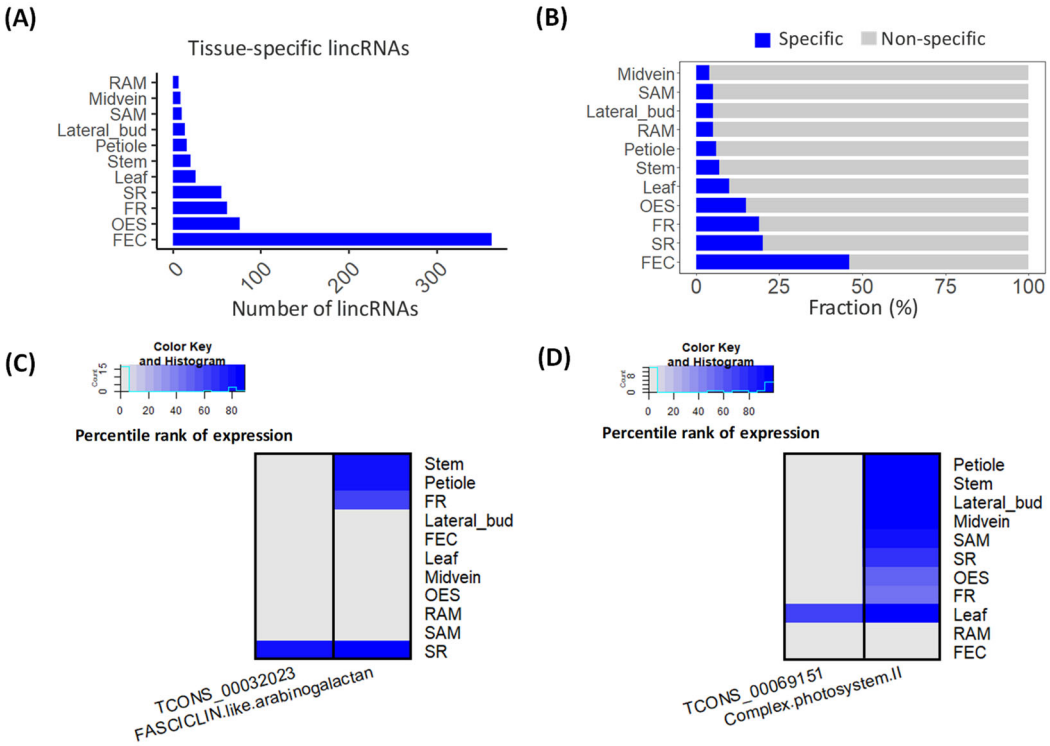

Figure 4 Potential functional lincRNAs in specific leaves or storage roots tissue. (A) The number of candidate tissue-specific lincRNAs in each cassava tissue. (B) The fraction of identified lincRNAs with specific and non-specific in 11 cassava tissues. (C) The expression level of potential SR specific lincRNA, TCONS_00032023 comparing to its *cis*-target gene encoding for FASCICLIN-like arabinogalactan. (D) The expression level of potential leaf specific lincRNA, TCONS_00069151 comparing to its *cis*-target gene encoding for Complex photosystem II

5. Conclusion

Here we provided a genome-wide collection of lincRNAs in 11 cassava tissues as common or tissue-specific lincRNAs. This is the first genome-wide analysis of lincRNAs under different cassava tissues including the shoot and root part of the cassava plant, and callus of cassava tissue culture. These putative lincRNAs might be a benefit for further functional validation and then revealing the complex post-transcription regulation in cassava, one of the most important crops for the 21st century.

6. Acknowledgments

The authors would like to thank the National Center for Genetic Engineering and Biotechnology (BIOTEC, NSTDA) for R.S post-graduate scholarship. We gratefully acknowledge Systems Biology and Bioinformatics (SBI), King Mongkut's University of Technology Thonburi (KMUTT) for computing facility.

7. References

Basu, D., Tian, L., Debrosse, T., Poirier, E., Emch, K., Herock, H., ... & Showalter, A. M. (2016). Glycosylation of a fasciclin-like arabinogalactan-protein (SOS5) mediates root growth and seed mucilage adherence via a cell wall receptor-like kinase (FEI1/FEI2) pathway in Arabidopsis. *PLoS ONE*, *11*(1), e0145092.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15-21.

Fatica, A., & Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics*, *15*(1), 7-21.

Golicz, A. A., Singh, M. B., & Bhalla, P. L. (2018). The Long Intergenic Noncoding RNA (LincRNA) Landscape of the Soybean Genome. *Plant Physiology*, *176*(3), 2133–2147.

Hao, Z., Fan, C., Cheng, T., Su, Y., Wei, Q., & Li, G. (2015). Genome-wide identification, characterization and evolutionary analysis of long intergenic noncoding RNAs in cucumber. *PLoS ONE*, *10*(3), 1-20.

Hu, M., Hu, W., Xia, Z., Zhou, X., & Wang, W. (2016). Validation of Reference Genes for Relative Quantitative Gene Expression Studies in Cassava (*Manihot esculenta* Crantz) by Using Quantitative Real-Time PCR. *Frontiers in plant science*, *7*, 1-12.

Huang, M. J., Zhao, J. Y., Xu, J. J., Li, J., Zhuang, Y. F., & Zhang, X. L. (2019). lncRNA ADAMTS9-AS2 Controls Human Mesenchymal Stem Cell Chondrogenic Differentiation and Functions as a ceRNA. *Molecular Therapy-Nucleic Acids*, *18*, 533-545.

Iempridee, T., Wiwithaphon, S., Piboonprai, K., Pratedrat, P., Khumkhrong, P., Japrung, D., Temisak, S., Laiwejpithaya, S., Chaopotong, P., & Dharakul, T. (2018). Identification of reference genes for circulating long noncoding RNA analysis in serum of cervical cancer patients. *FEBS open bio*, *8*(11), 1844–1854.

Jiang, N., Cui, J., Shi, Y., Yang, G., Zhou, X., Hou, X., ... & Luan, Y. (2019). Tomato lncRNA23468 functions as a competing endogenous RNA to modulate NBS-LRR genes by decoying miR482b in the tomato-Phytophthora infestans interaction. *Horticulture Research*, *6*(1), 1-11.

Khemka, N., Singh, V. K., Garg, R., & Jain, M. (2016). Genome-wide analysis of long intergenic non-coding RNAs in chickpea and their potential role in flower development. *Scientific reports*, *6*, 33297.

Kong, L., Zhang, Y., Ye, Z. Q., Liu, X. Q., Zhao, S. Q., Wei, L., & Gao, G. (2007). CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic acids research*, *35*(suppl_2), W345-W349.

Smid, M., van den Braak, R. R. C., van de Werken, H. J., van Riet, J., van Galen, A., de Weerd, V., ... & Wilting, S. M. (2018). Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data performs similarly in intersample analyses while improving intrasample comparisons. *BMC Bioinformatics*, *19*(1), 236.

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., Van Baren, M. J., ... & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515.

Varshney, D., Rawal, H. C., Dubey, H., Bandyopadhyay, T., Bera, B., Kumar, P. M., ... & Mondal, T. K. (2019). Tissue specific long non-coding RNAs are involved in aroma formation of black tea. *Industrial crops and products*, *133*, 79–89.

Wang, A., Hu, J., Gao, C., Chen, G., Wang, B., Lin, C., ... & Zhou, G. (2019a). Genome-wide analysis of long non-coding RNAs unveils the regulatory roles in the heat tolerance of Chinese cabbage (*Brassica rapa* ssp. chinensis). *Scientific reports*, *9*(1), 1-14.

Wang, L., Park, H. J., Dasari, S., Wang, S., Kocher, J. P., & Li, W. (2013). CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic acids research*, *41*(6), e74.

Wang, M., Zhao, W., Gao, L., & Zhao, L. (2018). Genome-wide profiling of long non-coding RNAs from tomato and a comparison with mRNAs associated with the regulation of fruit ripening. *BMC plant biology*, *18*(1), 75.

Wang, P., Dai, L., Ai, J., Wang, Y., & Ren, F. (2019b). Identification and functional prediction of cold-related long non-coding RNA (lncRNA) in grapevine. *Scientific reports*, *9*(1), 6638.

Wang, Y., Fan, X., Lin, F., He, G., Terzaghi, W., Zhu, D., & Deng, X. W. (2014). Arabidopsis noncoding RNA mediates control of photomorphogenesis by red light. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(28), 10359–10364.

Wilson, M. C., Mutka, A. M., Hummel, A. W., Berry, J., Chauhan, R. D., Vijayaraghavan, A., ... & Bart, R. S. (2017). Gene expression atlas for the food security crop cassava. *New Phytologist*, *213*(4), 1632-1641.

Yao, R. W., Wang, Y., & Chen, L. L. (2019). Cellular functions of long noncoding RNAs. *Nature Cell Biology*, *21*(5), 542–551.

Zhang, Y. C., Liao, J. Y., Li, Z. Y., Yu, Y., Zhang, J. P., Li, Q. F., ... & Chen, Y. Q. (2014). Genome-wide screening and functional analysis identify a large number of long noncoding RNAs involved in the sexual reproduction of rice. *Genome Biology*, *15*(12), 512.

Zhao, X., Gan, L., Yan, C., Li, C., Sun, Q., Wang, J., ... & Liu, J. N. (2019). Genome-Wide Identification and Characterization of Long Non-Coding RNAs in Peanut. *Genes*, *10*(7), 536.