Genomics 114 (2022) 110356

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Sequencing and *de novo* transcriptome assembly for discovering regulators of gene expression in Jack (*Artocarpus heterophyllus*)

Kishor U. Tribhuvan [a], Devendra K. Singh [a], Bhubaneshwar Pradhan [a], Sujit K. Bishi [a], Avinash Pandey [a], Sudhir Kumar [a], Jyotika Bhati [b], Dwijesh C. Mishra [b], Antra Das [c], T.R. Sharma [a], A. Pattanayak [a], Binay K. Singh [a,*]

[a] ICAR - Indian Institute of Agricultural Biotechnology, Ranchi 834 003, Jharkhand, India
[b] ICAR - Indian Agricultural Statistics Research Institute, New Delhi 110 012, India
[c] ICAR - Central Institute of Subtropical Horticulture, Lucknow 226 101, India

ARTICLE INFO

Keywords:
Artocarpus heterophyllus
Transcriptome assembly
lncRNA
eTM
Transcription regulators

ABSTRACT

Jack (*Artocarpus heterophyllus*) is a multipurpose fruit-tree species with minimal genomic resources. The study reports developing comprehensive transcriptome data containing 80,411 unigenes with an N50 value of 1265 bp. We predicted 64,215 CDSs from the unigenes and annotated and functionally categorized them into the biological process (23,230), molecular function (27,149), and cellular components (17,284). From 80,411 unigenes, we discovered 16,853 perfect SSRs with 192 distinct repeat motif types reiterating 4 to 22 times. Besides, we identified 2741 TFs from 69 TF families, 53 miRNAs from 19 conserved miRNA families, 25,953 potential lncRNAs, and placed three functional eTMs in different lncRNA-miRNA pairs. The regulatory networks involving genes, TFs, and miRNAs identified several regulatory and regulated nodes providing insight into miRNAs' gene associations and transcription factor-mediated regulation. The comparison of expression patterns of some selected miRNAs *vis-à-vis* their corresponding target genes showed an inverse relationship indicating the possible miRNA-mediated regulation of the genes.

1. Introduction

Jack (*A. heterophyllus*) (2n = 4× = 56) is a member of the family Moraceae [1]. It is native to the western coast of India. However, today it grows throughout tropical and subtropical regions of the world, mainly along the roadsides and occasionally in forests and woodlots [2]. Jack is an evergreen, outcrossing monoecious tree with unisexual flowers appearing separately on its trunk and the aged branches. It bears small (~4 mm across) flowers clustered in an inflorescence. The stamen and ovary present singly in the male and female flowers are surrounded by a single whorl of perianth tissues. On pollination, a female flower develops into a fruit with the perianth tissue swelling into sweet flesh. Thousands of flowers in the female inflorescence form a syncarp, resulting in the largest tree-borne fruit structure, sometimes reaching 50 kg [3].

The fruit is the principal economic part of Jack; all the other parts and their components also have significant commercial or nutritive and non-nutritive value [4]. For example, latex obtained from Jack contains serine proteases useful for milk clotting and meat tenderization in dairy and food processing industries [5], and mature seeds yield carbohydrate and protein-rich flour useful for preparing gluten-free bakery products and dough thickening [6–10]. Seed flour also finds applications in preparing various other food products, including jam, wine, fermented beverages, and ice cream [11]. Like several other less known fruits growing naturally across different climatic regions, Jack fruits are rich in bioactive compounds such as flavonoids, phenolics, anthocyanins, and nutritional compounds such as sugars, carotenoids, vitamins, and minerals. Besides, they have a distinct flavour and taste. The bioactive compounds of Jack fruits are significant to human health and are potential drug candidates [4,11–15].

Categorized initially as an underutilized fruit tree crop, Jack is increasingly gaining global popularity due to its increasing demand by Asian immigrants as meat substitutes. Its potential health value to humans and expanding food and non-food domestic and industrial applications further increase its popularity [16]. However, despite its vast significance, this tropical fruit tree crop is minimally researched and

* Corresponding author.
E-mail address: binay.singh@icar.gov.in (B.K. Singh).

https://doi.org/10.1016/j.ygeno.2022.110356
Received 23 November 2021; Received in revised form 12 March 2022; Accepted 27 March 2022
Available online 29 March 2022

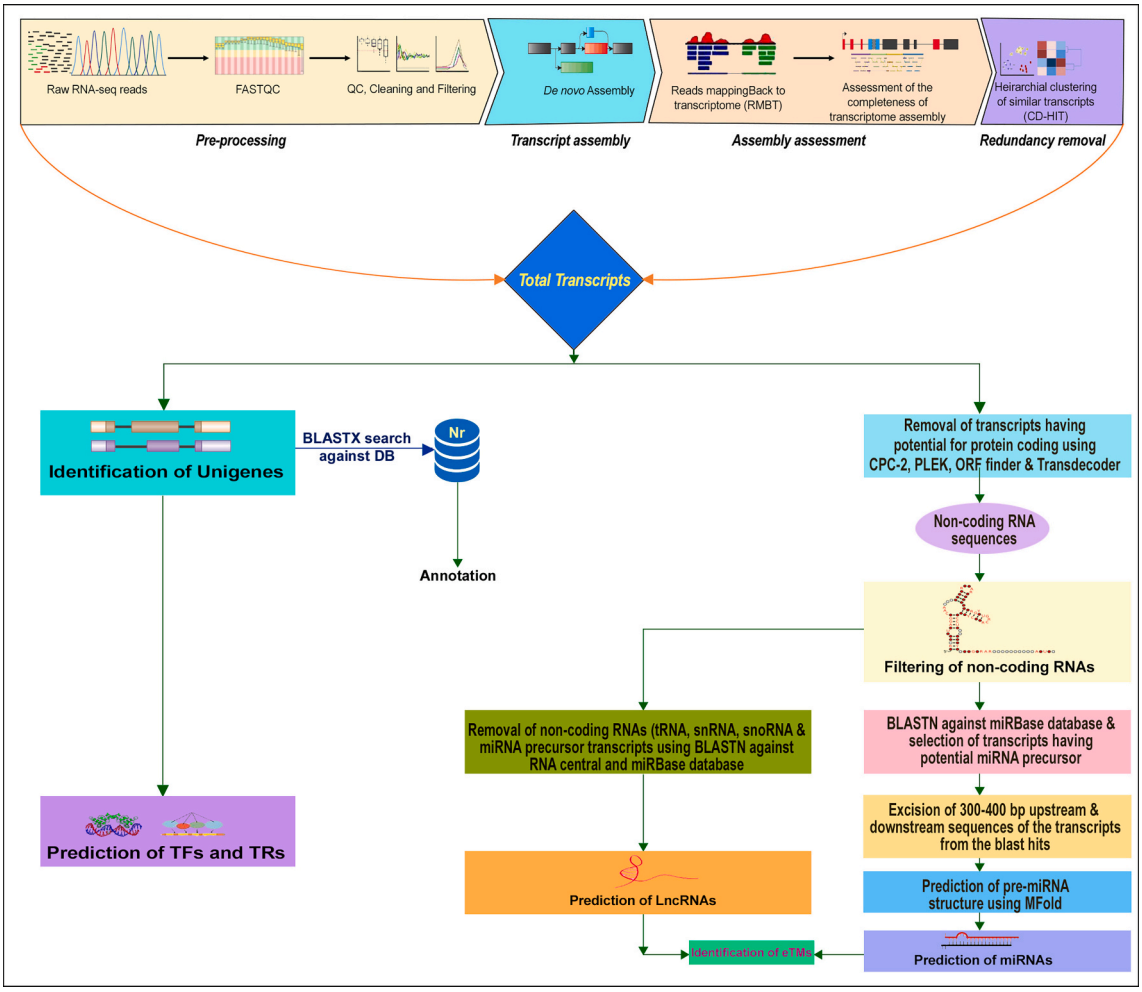

Fig. 1. Schematic representation of RNA-seq data analysis workflow.

needs extensive genetic improvement to meet the end-user needs and preferences [17].

Conventionally, the genetic improvement in trees involves selecting the desirable trees followed by mating and testing the resulting progeny. The fundamental limitation of this strategy is that improving one trait often accompanies the appearance of one or a few undesirable characters [18]. We can address this issue by employing trait-linked markers and discovering and utilizing the regulators of gene expression affecting the traits of interest. Regarding the latter, microRNAs (miRNAs), which are among the most critical riboregulators that fine-tune the expression of diverse regulatory genes affecting growth, development, and stress responses in plants and trees, are highly promising [19].

During the last few years, plant scientists have developed several miRNA-based approaches for crop improvement. Crop improvement strategies employing positively regulated miRNAs often involve their constitutive overexpression. In the case of miRNAs functioning as negative regulators, downregulation or loss-of-function of the miRNAs, overexpression of miRNA-resistant form of the target or artificial miRNA–target mimic are effective strategies [20]. These strategies have been extensively employed to analyze miRNAs and miRNA-based trait improvement in several crop plants [21]. The success in crop plants indicates that establishing miRNAs in a scientifically under-explored tree species, like Artocarpus, would help better understand the fundamentals of growth and development and its ability to tolerate different stresses. Moreover, it would enhance our capability to effectively manipulate various traits to improve the utility of the crop. Regarding these, the progress in Jack has been minimal. But the recently published draft genome sequence of Jack has made gene/pathway-targeted surveys possible to some extent. Some critical fruit-related pathways, including starch and sucrose metabolism, have been preliminarily investigated [22,23]. Further genomic coupled with multi-omics studies would make whole-genome and gene-targeted surveys possible to a more considerable extent [24,25].

We report developing a comprehensive transcriptome dataset, annotation and functional classification of unigenes, mining and *in-silico* characterization of simple sequence repeats (SSRs), identifying transcription factors (TFs) and transcription regulators (TRs), discovering miRNAs and long non-coding RNAs (lncRNAs), and identifying their targets and functional annotation in Jack. Moreover, we predicted endogenous target mimics (eTMs) and validated a set of randomly selected miRNA and their targets through expression analysis. The results obtained here will provide valuable resources for more extensive molecular studies in Jack. Moreover, it will lay a scientific foundation for designing efficient genetic improvement and conservation strategies in the crop.

2. Material and methods

2.1. RNA extraction and cDNA library preparation

Total RNA was isolated separately from one gram of developing seeds, leaves, inflorescence, and the roots of the *A. heterophyllus* (Acc. No. IC436479) using ZR Plant RNA MiniPrep Kit (Zymo Research, CA, USA). The RNA was qualitatively analyzed on 1.0% denaturing agarose

gel and quantified on NanoDrop 1000 (NanoDrop Technologies, USA). One μg of total RNA extracted from each of the four tissue types was mixed to prepare an RNA pool. The quality-checked RNA pool served as the substrate to prepare an RNA-seq paired-end sequencing library using Illumina TruSeq Stranded mRNA Library Preparation Kit (Illumina, San Diego, USA). The quality check of the library was performed in an Agilent 4200 TapeStation System (Agilent Technologies, CA, USA) using high sensitivity D5000 Screen Tape. The library was quantitatively analyzed on Qubit Fluorometer using Qubit dsDNA BR Assay Kit (ThermoFisher Scientific, USA).

2.2. Sequencing and sequence assembly

We used an Illumina NextSeq1000 platform to sequence the library. The data analysis pipeline is depicted in Fig. 1. The quality of the raw data obtained by paired-end sequencing of the normalized cDNA library was assessed using FastQC software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). The adaptor contamination and low-quality reads were removed using Trimmomatic v0.32 software [26]. The study included only those sequence reads whose lengths were more than 50 nucleotides. We cleaned the sequences thoroughly by removing ambiguous reads (unknown nucleotides 'N' > 5%), low-quality reads (QV < 20), and adaptor sequences. The clean reads were assembled *de novo* using Trinity software v2.11.0 [27] with the default *k-mer,* K = *25,* and redundancies in the assembled transcript sequences were removed using *C*D-HIT-EST v4.6 software [28] with global sequence identity threshold cut-off set at 90%. The unigenes were identified using Perl script"get_longest_isoform_seq_per_trinity_gene.pl".

2.3. Mapping of reads to transcriptome assembly

We used Bowtie2 v2.3.5.1 and Burrows-Wheeler Alignment (BWA) v0.7.12 software [29] for mapping the reads back to the transcriptome assembly. The 'end-to-end' and the 'sensitive' options were used for the alignment using Bowtie2, in which the max number of mismatches (*-N* = 1) was allowed in the seed alignment. For BWA, 'bwtsw' algorithm was used to index the assembled transcriptome data, and the bwa-mem with default parameters was used to align the reads. SAMtools v1.7 was used to calculate mapping statistics from BAM files [30].

2.4. Benchmarking universal single-copy orthologs (BUSCO) analysis

We used Benchmarking Universal Single-Copy Orthologs (BUSCO) software v2.0 to evaluate the quality and completeness of the *de novo* assembly. The analysis was done using the transcriptome assessment mode with the eukaryote lineage database (eukaryota_orthoDB9) and viridiplantae lineage database (viridiplantae_orthoDB10).

2.5. Coding DNA sequences (CDS) prediction and functional annotation

We used TransDecoder software v5.0.2 to predict coding sequences from the unigenes. The predicted coding sequences were compared against NCBI non-redundant protein database (nr) using BLASTx with an e-value threshold of $1e^{-5}$ and assigned functions based on sequence similarity to proteins of known functions. We used Blast2GO software [31] preloaded with local nr database to get gene ontology (GO) annotation defined by the biological process, molecular function, and cellular component; GO functional classification; predict and classify their probable functions, and mapping to reference canonical pathways in the KEGG database.

2.6. Identification of SSRs, TFs, and TRs

SSR mining was performed using MISA (microsatellite search module) with 2–6 bp repeats described by [32]. TFs and TRs were identified with the help of the PlantTFcat online tool (http:// plantgrn.noble.org/PlantTFcat).

2.7. Identification of miRNA, lncRNA, and their targets

We used the set of non-coding unigenes for identifying the miRNAs and lncRNAs. The non-coding unigenes were separated from the potential protein-coding unigenes using the coding-potential prediction tools, ORFfinder (http://www.ncbi.nlm.nih.gov/gorf/orfig.cgi), TransDecoder (v2.1.0) (https://transdecoder.github.io/), Coding Potential Calculator (CPC2), and predictor of long non-coding RNAs and messenger RNAs based on an improved *k-mer* scheme (PLEK v1.2). The potential pre-miRNA sequences were identified by using the non-coding unigenes as queries for BLASTn searches against the miRNA repositories, miRbase release 22.1 and plant microRNA database (PMRD) using the Unipro UGENE v38 [33] following the criteria: mismatch less than 3, word size 7 and e-value $\leq 1e^{-5}$. The query sequences with a significant match (80% sequence similarity and 100% query coverage) in one or both the miRNA repositories were selected and used to deduce a non-redundant set of potential pre-miRNAs. We further analyzed the miRNA precursor sequences comprising 300–400 nucleotides flanking the 3′ and 5′ ends of the potential miRNAs in the pre-miRNA sequences to RNA folding using Mfold (http://mfold.rna.albany.edu/?q= mfold) set at default parameters [34]. The sequences satisfying the miRNA annotation criteria: perfect stem-loop hairpin, Minimal Folding Free Energy Index (MFEI) $\geq$ 0.41, AU content between 22 and 77%, no loop or break in miRNA* sequence, miRNA* with less than 6 mismatches, mature miRNAs at one arm of the hairpin, SSR signature value $R \geq 2.5$, Normalized Shannon Entropy (NQ) $\leq$ 0.45, normalized base-pair distance (ND) $\leq$ 0.15, and normalized base-pairing propensity (Npb) $\geq$ 0.25, were considered potential miRNA sequences [35,36]. The potential targets of the predicted miRNAs were identified by searching the miRNA sequences against the *de novo* assembled unigenes using the psRNATarget server [37].

For identifying lncRNAs, the non-coding unigenes were subjected to removing all the miRNA and small RNA, including small nuclear (snRNA), small nucleolar RNAs (snoRNA), transfer RNA (tRNA), small interfering RNA (siRNA), and ribosomal RNA (rRNA) precursor transcript sequences. The precursor transcript sequences for small RNAs were identified using the non-coding unigenes as queries for BLASTn searches against the RNAcentral database. We considered the remaining unigenes as potential lncRNAs and identified their targets by comparing them with the *de novo* assembled unigenes using BLASTn with a cut-off e-value $\leq$ 1e−5 and $\geq$ 95% identity. The targets for the potential lncRNAs were identified based on the binding energy of the lncRNA-target complex analyzed using RNAplex v2.4.18 software. The transcripts participating in the lncRNA-mRNA complex with binding energy (ΔG) lower than −50 kcal/mol were considered potential targets.

2.8. TF-target networking

We predicted the targets of the TFs by analyzing their nucleotide sequences using the psRNAtarget tool, considering *A. thaliana* as the reference. We deduced the regulatory connectivity between the TFs and their targets by analyzing the small RNA targets associated with TFs and their target accessibility value using Cytoscape software [38]. The maximum cut-off score was kept below 2.5 for analysis.

2.9. Experimental validation of miRNA and miRNA-target interaction

We experimentally validated the miRNAs and their interaction with mRNA targets by the quantitative real-time PCR (qRT-PCR) expression analysis of a set of selected miRNAs and their target genes involved in leaf development and hormone signaling. The primers for the qRT-PCR analysis were designed using the IDT oligo analyzer tool (https://www.idtdna.com/calc/Analyzer) and the Minimum Information for Publication of Quantitative Real-Time PCR Experiments (MIQE) guidelines

Table 1
Summary of transcriptome *de novo* assembly in *Artocarpus heterophyllus*.

S. No.	Particulars	Transcripts	Unigenes
1	Number of sequences	1,39,384	80,411
2	Average length (bp)	1002	682
3	N50 (bp)	1723	1265
4	Minimal length (bp)	200	200
5	Maximal length (bp)	9601	9601
6	Median length (bp)	592	345
7	Total assembled bases	139,662,768	54,840,302
8	GC content	41.94%	41.68%

[39]. Total RNA was isolated from the leaves of *A. heterophyllus* at four different stages, namely leaf primordia, tender leaf, young leaf, and mature leaf using miRNAeasy kit. The RNAs isolated from two biological replicates were pooled for gene expression analysis experiments. For mRNA expression analysis, approximately one μg of DNaseI treated total RNA was used for the first-strand cDNA synthesis using QuantiTect reverse transcriptase kit (Qiagen, Cat No. 205311). While 300 ng of total RNA was used to synthesize cDNA from miRNA using miScript-plant RT kit (Qiagen, Cat No. 218761) for miRNA expression analysis. To quantitatively validate the predicted target transcripts and miRNAs, we performed qRT PCR expression analysis using SYBR Select Master Mix (Thermo Fisher Scientific, Cat No. 4472908) and miScript SYBR Green PCR Kit (Qiagen, Cat No. 218073), respectively. The primers used in the experiment are listed in Supplementary Table S1. We used the Jac α-tubulin gene and the U6 snRNA gene as endogenous reference genes for the target and miRNA expression studies.

2.10. eTM prediction

We predicted the eTMs based on the pairing between lncRNA and miRNA analyzed using psRobot software. The following parameters were considered for the prediction of eTMs: (1) bulges should be present only at the 5′ end of the miRNA sequence extending between 9th to 12th positions; (2) only three nucleotides bulge in eTMs was permitted; (3) perfect nucleotide pairing at the 5′ end of miRNA between 2nd to 8th positions; and (4) except for the central bulge, the total number of mismatches and G/U pairs within eTM and miRNA pairing regions should not be more than three.

3. Results

3.1. Sequencing of cDNA library, and transcriptome de novo assembly and evaluation statistics

We used the Illumina NextSeq1000 platform for paired-end sequencing of the RNAseq library and generated 42,928,887 paired-end raw reads. We have deposited the raw read sequence data at the NCBI, SRA; Acc. No. SRR7250836. Cleaning the raw reads for adaptors and low-quality reads using Trimmomatic software v0.32 yielded 41,549,555 clean reads. The *de novo* assembly of the clean-reads using Trinity software v2.11.0 generated 1,39,384 contigs (transcripts hereafter) containing 139,662,768 nucleotides. A total of 80,411 unigenes containing 54,840,302 nucleotides were extracted from these transcripts using a Perl script get_longest_isoform_seq_ per_trinity_gene.pl in Trinity. The N50 values for transcripts and unigenes were 1723 bp and 1265 bp, respectively. The lengths of the transcripts and unigenes ranged between 200 and 9601 bp. The average lengths of transcripts and unigenes were 1002 and 682 bp, while the median lengths were 592 and

Table 2A
Bowtie2 and BWA alignment statistics of cleaned reads to the *de novo* transcriptome assembly in *Artocarpus heterophyllus*.

Particulars	Bowtie2	BWA
Total paired-end reads	41,521,099	41,847,494
Reads aligned 01 time	40,311,955 (97.08%)	39,449,805 (94.27%)
Reads aligned >1 times	0 (0%)	26 (0.00006%)
Reads not aligned	1,209,144 (2.91%)	2,397,663 (5.7%)
Overall alignment rate (%)	97.08%	94.27%

Table 2B
BUSCO analysis for assessing transcriptome assembly completeness with the eukaryote lineage database (eukaryota_orthoDB9) and viridiplantae lineage database (viridiplantae_orthoDB10).

	Number of BUSCO units found	
	Eukaryota_orthoDB9	Viridiplantae_orthoDB10
Complete BUSCOs	225 (88.2%)	329 (77.4%)
Complete and single-copy BUSCOs	216 (84.7%)	324 (76.2%)
Complete and duplicated BUSCOs	9 (3.5%)	5 (1.2%)
Fragmented BUSCOs	19 (7.5%)	70 (16.5%)
Missing BUSCOs	11 (4.3%)	26 (6.1%)
Total BUSCOs searched	**255 (100%)**	**425 (100%)**

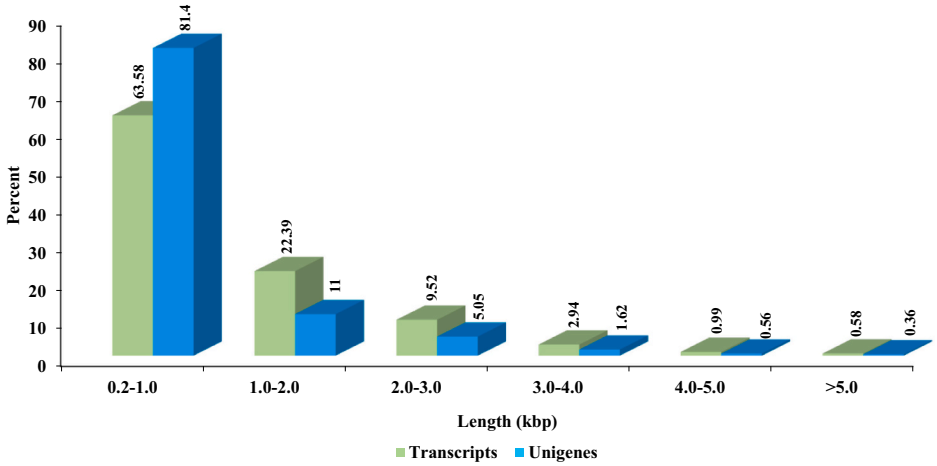

Fig. 2. Length-wise distribution frequency of assembled transcripts and unigenes.

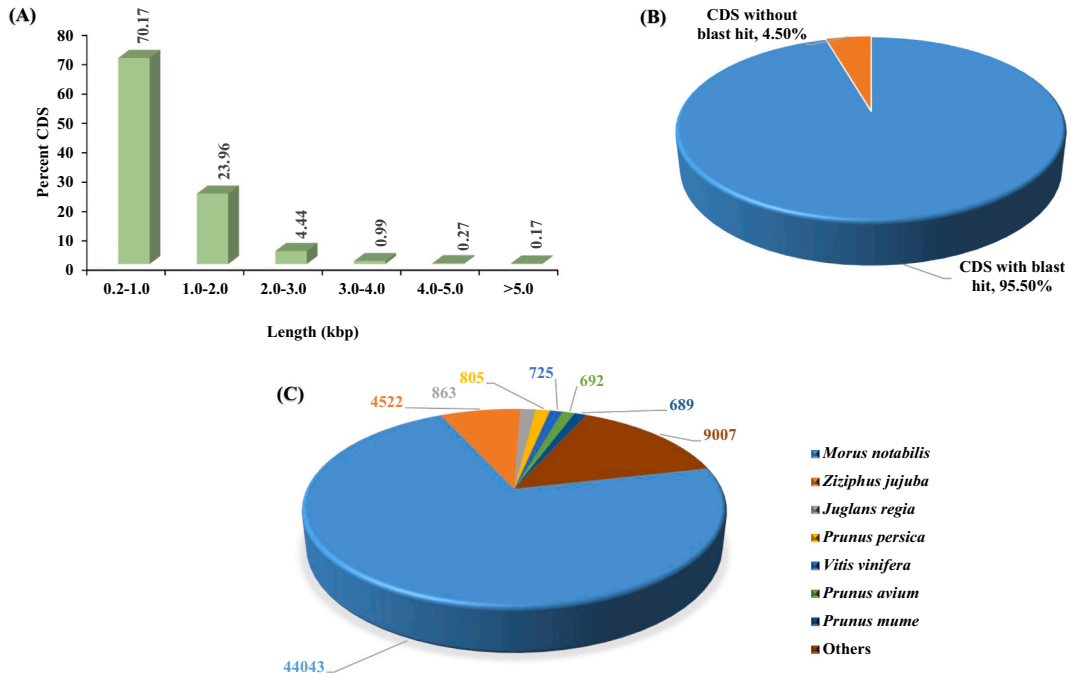

Fig. 3. (A) Length-wise distribution frequency of CDSs, (B) Distribution of CDSs with or without blast hit against nr database, (C) Homology-based species distribution of CDSs.

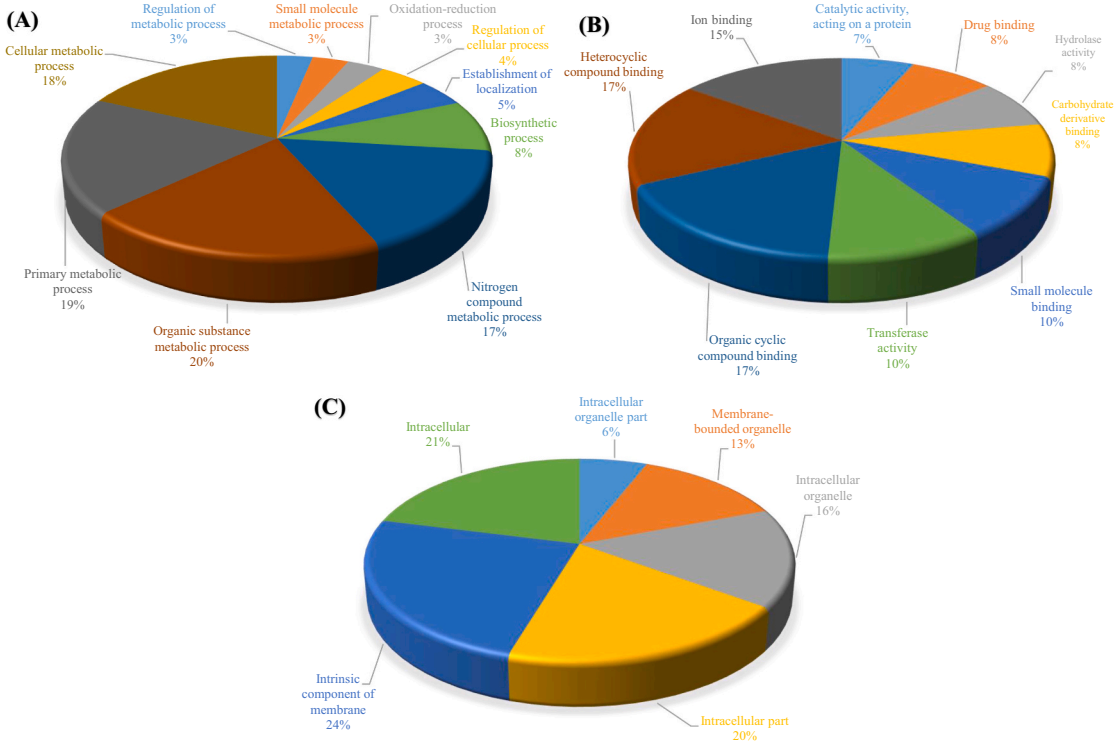

Fig. 4. Functional classification of CDSs assigned with GO terms, (A) Biological process, (B) Molecular function, (C) Cellular component.

345 bp. The GC contents of transcripts and unigenes were 41.94% and 41,68%, respectively. The summary of the transcriptome *de novo* assembly is indicated in Table 1. A total of 63.58% transcripts were in the range of 0.2–1.0 kbp, 22.39% in 1.0–2.0 kbp, 9.52% in 2.0–3.0 kbp, 2.94% in 3.0–4.0 kbp, 0.99% in 4.0–5.0 kbp and 0.58% >5.0 kbp, while these counts for the unigenes were 81.4%, 11.0%, 5.05%, 1.62%, 0.56% and 0.36% respectively (Fig. 2). Since there is no reference sequence for Jack, we considered the *de novo* assembled transcriptome sequence as the reference sequence to evaluate the quality of the transcripts. Overall, 97.08 and 94.27% of the clean reads successfully mapped back to the reference sequence using Bowtie2 and BWA alignment tools. The statistics of cleaned reads mapping back to the *de novo* assembled transcriptome sequence is summarized in Table 2A. We made the quantitative assessment of the completeness of the transcriptome using

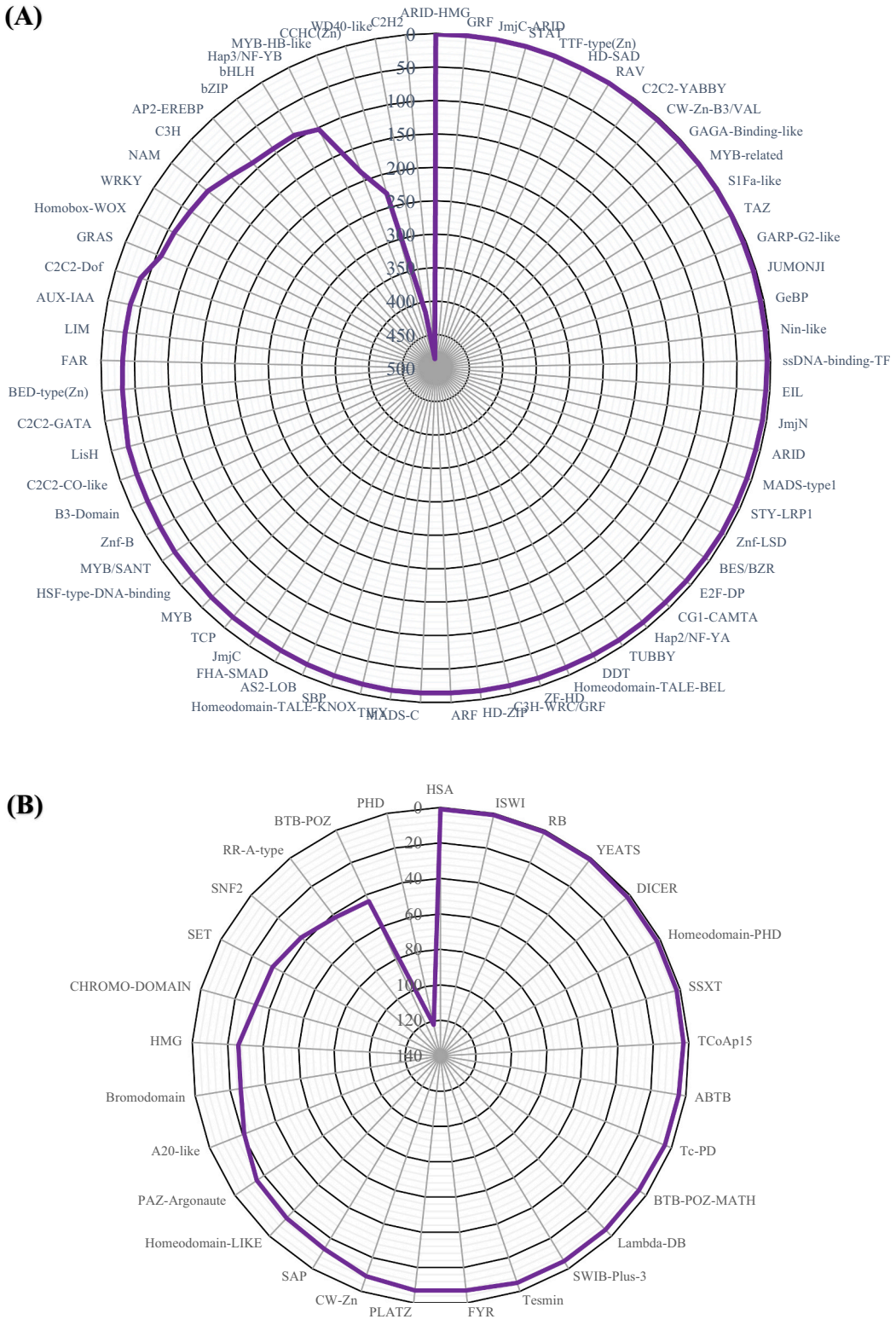

Fig. 5. (A) Transcription factors and (B) Transcription regulators identified in *Artocarpus heterophyllus* transcriptome.

eukaryota and viridiplantae BUSCO databases. The results obtained through BUSCO analysis using eukaryota and viridiplantae databases is summarized as C**:** 88.2% [S: 84.7%, D: 3.5%], F: 7.5%, M: 4.3%, n: 255 and C**:** 77.4% [S: 76.2%, D: 1.2%], F: 16.5%, M: 6.1%, n: 425, respectively where C = complete, S = complete and single-copy, D = complete and duplicated, F = fragmented, M = missing, and n = total number of BUSCOs identified (Table 2B).

3.2. Prediction and annotation of CDS

Using the TransDecoder software v5.0.2, we predicted 64,215 CDSs from 80,411 unigenes. A total of 70.17% CDSs were in the range of 0.2–1.0 kbp, 23.96% in 1.0–2.0 kbp, 4.44% in 2.0–3.0 kbp, 0.99% in 3.0–4.0 kbp, 0.27% in 4.0–5.0 kbp and 0.17% >5.0 kbp (Fig. 3A). Among the predicted CDSs, 61,346 (95.5%) CDSs returned a positive

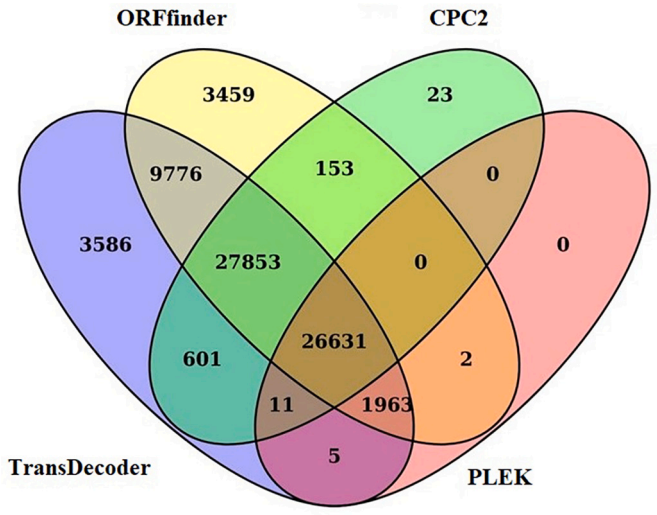

Fig. 6. VENN diagram representing the number of non-coding unigenes identified through different computational tools.

BLAST hit with known unigenes available in the nr database, while 2869 (4.5%) CDSs could not find any hit and thus considered unique to Jack (Fig. 3B). Most BLAST top hits were *Morus notabilis* (44,043; 71.8%). The other tree/plant species showing significant hits were *Ziziphus jujube* (4522; 7.4%), *Juglans regia* (863; 1.4%), and *Prunus persica* (805; 1.3%) (Fig. 3C). Based on GO annotation using Blast2GO pipeline, we assigned GO terms to 34,231 (53.3%) CDSs and classified them into three main categories: biological process (23,230), molecular function (27,149), and cellular components (17,284). In the biological process category, organic substance metabolic process (20%), primary metabolic process (19%), cellular metabolic process (18%), and nitrogen compound metabolic process (17%) were the predominant subcategories (Fig. 4A), while in the molecular function category, heterocyclic compound binding (17%), organic cyclic compound binding protein (17%) and ion binding (15%) (Fig. 4B), and in the cellular component category, the intrinsic component of membrane (24%) followed by intracellular (21%) and intracellular part (20%) were the predominant subcategories (Fig. 4C).

3.3. Identification and in-silico characterization of SSRs

We identified 16,853 perfect SSRs in 15,012 of 80,411unigenes, accounting for one SSR per 6.4 kbp of unigene sequence. Among the 16,853 SSR loci analyzed, trinucleotides were the most prevalent (80.57%). The frequencies of di-, tetra-, penta- and hexanucleotide repeats were 14.70, 3.04, 1.1, and 0.62%. Twelve bp SSRs were most frequent (48.26%) followed by 15 bp (17.11%), 18 bp (7.53%) and 20 bp (6.66%). GAA/TTC and AG/CT were the most abundant tri- and dinucleotide repeat motifs with 10.67 and 4.98% frequencies. The frequency of the repeat motifs varied from 48.26% ($n = 4$) to less than 1% ($n > 12$). A total of 192 distinct repeat motif types were identified in the study. Different repeat motifs were reiterated 4 to 22 times (Supplementary Fig. 1A, B & C).

3.4. Identification of TFs and TRs

We used the PlantTFcat online tool to identify TFs and TRs from the transcriptome data. Of the 80,411 unigenes, we cataloged 2741 unigenes into 69 TF families and 489 unigenes into 29 TR families. Among the 69 TF families, C2H2 (486) was most abundant, followed by WD40-like (415), CCHC (zn) (229), MYB-HB-like (187), Hap3/NF-YB (103), bZIP (92), and WRKY (65) (Fig. 5A), while among the 29 TR families, PHD (122) was most abundant followed by BTB-POZ (44), RR-A-type (42), SNF2 (37), SET (33), and Chromo-Domain (32) (Fig. 5B).

3.5. Identification of miRNAs and their targets genes

Screening through ORFfinder, TransDecoder, CPC2, and PLEK software recognized 69,837; 70,426; 55,272; and 28,612 non-coding unigenes, respectively, among the 80,411 assembled unigenes. A total of 26,631 non-coding unigenes only were, however, identified commonly by all the four computational approaches (Fig. 6). BLASTn homology search for 26,631 non-coding unigenes against non-redundant plant miRNA sequences available at miRbase and PMRD identified 360 potential pre-miRNAs. The analysis of miRNA precursor sequences comprising 300–400 nucleotides flanking the 3′ and 5′ ends of the putative miRNAs in the pre-miRNA sequences using Mfold software revealed that 53 miRNA precursors could form miRNA-like-foldback structures. The miRNA target prediction analysis using psRNATarget online web server identified 31 unigenes as targets for 53 miRNAs with the expectation values ranging from 0 to 3.5. The analyzed miRNA sequences and their corresponding target unigenes are detailed in Table 3. The potential pre-miRNA sequences are detailed in Supplementary Text 1. The secondary structures of precursor sequences of selected miRNAs are depicted in Fig. 7. Family-wise distribution of the miRNAs revealed that these miRNAs belonged to 19 conserved miRNA families. The most abundant were conserved miRNAs from miR166 (11), miR172 (10), miR396 (6), and miR156 (6). The targets for these miRNAs mainly were TFs involved in various developmental pathways.

3.6. TF-target regulatory network

TFs regulate the target genes directly or interact with miRNA, affecting the target gene's expression. The gene-TFs-miRNAs regulatory network identified several potential regulatory and regulated nodes and provided insight into the gene associations, transcription factors-mediated regulation, and control by miRNAs. We identified a total of 200 nodes and 484 edges after filtering the score and obtained the maximum gene-TFs–miRNA interaction for SBP followed by C2H2, MYB, and Hap2/NF-YA transcription factors (Fig. 8). The most targeted miRNAs were miR5658, miR5021, miR414, followed by miR858.

3.7. Identification of lncRNAs and their target genes

BLASTn homology search for 26,271 non-coding unigenes, remaining after filtering out 360 pre-miRNA sequences, against the RNAcentral database further removed 318 unigenes significantly matching snRNA, snoRNA, tRNA, siRNA, and rRNA sequences. Thus a total of 25,953 unigenes were considered potential lncRNAs by all methods. lncRNA target prediction analysis identified a total of 5350 targets for the lncRNAs (Supplementary Table S2). Based on the pairing analysis between 53 miRNAs and the entire set of 25,953 potential lncRNAs using psRobot software, we placed three functional eTMs in the lncRNA-miRNA pairs: DN45875- ahe-miR004, DN11417- ahe-miR010, and DN67158-ahe-miR038. The targets of the miRNAs ahe-miR004, ahe-miR010, and ahe-miR038, namely DN14178 (unannotated), DN3518 (coding for extra-large guanine nucleotide-binding protein), DN8925 (coding for postsynaptic protein), respectively, were considered as the potential targets for the identified eTMs.

3.8. Expression analyses of miRNAs and their target genes

We compared the expression patterns of seven miRNAs *vis-à-vis* their corresponding target genes at four different stages of leaf development in *A. heterophyllus*. Of the seven miRNAs, the expression patterns of three miRNAs, namely ahe-miR008, ahe-miR013, ahe-miR050, indicated an inverse relationship with the expression patterns of target genes DN4470, DN934 and DN5647, respectively (Fig. 9), indicating their possible miRNA-mediated regulation. All three miRNAs showed

Table 3
Predicted miRNAs and their target transcripts in *Artocarpus hetrophyllus*.

Predicted_miR_ID	miRNA family	Predicted miRNA seq_RNA form	Nucleotide length	Precursor miRNA Transcript_ID	Target transcript ID	Expectation value	Description	eTMs
ahe-miR001	miR156	UUGACAGAAGAUAGAGAGC	19	DN8406_c0_g3_i1	DN5506_c0_g1_i12	1.5	squamosa promoter-binding-like protein 2	—NA—
ahe-miR002	miR156	AGAAAGAGAAGUGAGCACGCAC	22	DN13519_c0_g1_i3	DN19841_c0_g1_i1	1.5	—NA—	—NA—
ahe-miR003	miR156	CAGAAGAUAGAGAGCACAAC	20	DN8406_c0_g3_i1	DN36251_c0_g1_i1	1.5	squamosa promoter-binding-like protein 13A	—NA—
ahe-miR004	miR156	CUUUGAAAGUAGUAAAAGCCCU	22	DN21215_c0_g1_i1	DN14178_c0_g1_i1	2	—NA—	DN45875_c0_g1_i1
ahe-miR005	miR156	CUCUCUAUCUUCUGUCAACAUU	22	DN8406_c0_g3_i1	DN1149_c1_g1_i1	1.5	galactose-binding domain-like	—NA—
ahe-miR006	miR156	UGACAGAAGAUAGAGAGCAC	20	DN8406_c0_g3_i1	DN5506_c0_g1_i12	1.5	squamosa promoter-binding-like protein 2	—NA—
ahe-miR007	miR159	GUUGAGGUGAUUAAUAAAUAUU	22	DN34963_c0_g1_i1	DN14861_c0_g1_i2	1.5	probable leucine-rich repeat receptor-like protein kinase At2g33170	—NA—
ahe-miR008	miR160	GCCUGGCUCCCUGUAUGCCAU	21	DN1695_c1_g1_i3	DN4470_c0_g1_i7	0.5	auxin response factor	—NA—
ahe-miR009	miR160	UGCCUGGCUCCCUGUAUGCCAUU	23	DN1695_c1_g1_i3	DN4470_c0_g1_i7	0.5	auxin response factor	—NA—
ahe-miR010	miR162	GAUGAGAGAGAGAGAGAGAG	22	DN8885_c0_g2_i1	DN3518_c0_g2_i1	0	extra-large guanine nucleotide-binding protein 1	DN11417_c0_g1_i1
ahe-miR011	miR162	AGUGAGCGCUGGAUGCAGAGGU	22	DN1149_c3_g2_i1	DN23352_c0_g1_i1	2.5	—NA—	—NA—
ahe-miR012	miR162	AAUCUUUCUUCCUUACUUUUUU	22	DN1149_c3_g3_i1	DN66745_c0_g1_i1	2	—NA—	—NA—
ahe-miR013	miR166	UCGGACCAGGCUUCAUUCCUC	21	DN11264_c0_g2_i1	DN934_c0_g1_i2	0.5	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR014	miR166	UCGGACCAGGCUUCAUUCC	19	DN11264_c0_g2_i1	DN934_c0_g1_i2	0.5	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR015	miR166	UCGGACCAGGCUUCAUUCCCCC	22	DN1623_c1_g1_i3	DN934_c0_g1_i2	0.5	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR016	miR166	GGAAUGUUGUCUGGCUCGAGG	21	DN1623_c1_g1_i3	DN7276_c0_g1_i3	3	serine/threonine-protein kinase GRIK1	—NA—
ahe-miR017	miR166	UCGGACCAGGCUUCAUUCCCUU	22	DN1623_c1_g1_i3	DN934_c0_g1_i2	0.5	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR018	miR166	UCGGACCAGGCUUCAUUCCU	20	DN11264_c0_g2_i1	DN934_c0_g1_i2	0.5	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR019	miR166	UCGGACCAGGCUUCAUUCCCCU	22	DN1623_c1_g1_i3	DN934_c0_g1_i2	0.5	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR020	miR166	UCGGACCAGGCUUCAUUCCCG	21	DN1623_c1_g1_i3	DN934_c0_g1_i2	0.5	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR021	miR166	GGACCAGGCUUCAUUCCCC	19	DN1623_c1_g1_i3	DN934_c0_g1_i2	2	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR022	miR166	UCGGACCAGGCUUCAUUCCC	20	DN1623_c1_g1_i3	DN934_c0_g1_i2	0.5	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR023	miR166	CGGACCAGGCUUCAUUCCCC	20	DN1623_c1_g1_i3	DN934_c0_g1_i2	1	homeobox-leucine zipper protein ATHB-15	—NA—
ahe-miR024	miR167	CACCAACAAUAGAAGAACUUCA	22	DN18421_c0_g2_i1	DN13011_c0_g2_i1	2	protein HEADING DATE 3B-like isoform X1	—NA—
ahe-miR025	miR167	ACCGUGCACCACCAGCAGUUGA	22	DN18366_c0_g1_i2	DN20423_c0_g1_i1	3.5	—NA—	—NA—
ahe-miR026	miR168	CAAGCGAAUUAGAGACCGCCGG	22	DN7549_c0_g1_i3	DN51632_c0_g1_i1	3.5	U-box domain-containing protein 11	—NA—
ahe-miR027	miR169	AGAGGUAGAGAUUUGAAUGCAG	22	DN41965_c0_g1_i1	DN37441_c0_g1_i1	2	—NA—	—NA—
ahe-miR028	miR172	UGAGAAUCUUGAUGAUGCUGCAU	23	DN8633_c1_g1_i1	DN38291_c0_g1_i1	1	pentatricopeptide repeat-containing protein At1g20300, mitochondrial	—NA—
ahe-miR029	miR172	GGGAAUCUUGAUGAUGCUGCA	21	DN33154_c0_g1_i2	DN35094_c0_g1_i20	0.5	AP2-like ethylene-responsive transcription factor TOE3 isoform X1	—NA—
ahe-miR030	miR172	AGAAUCUUGAUGAUGCUGCAU	21	DN8633_c1_g1_i1	DN56487_c0_g1_i1	1	—NA—	—NA—
ahe-miR031	miR172	GGAAUCUUGAUGAUGCUGCA	20	DN33154_c0_g1_i2	DN35094_c0_g1_i20	0.5	AP2-like ethylene-responsive transcription factor TOE3 isoform X1	—NA—
ahe-miR032	miR172	GGAAUCUUGAUGAUGCUGCAU	21	DN33154_c0_g1_i2	DN35094_c0_g1_i20	0.5	AP2-like ethylene-responsive transcription factor TOE3 isoform X1	—NA—
ahe-miR033	miR172	UGAGAAUCUUGAUGAUGCUGC	21	DN8633_c1_g1_i1	DN38291_c0_g1_i1	1	pentatricopeptide repeat-containing protein At1g20300, mitochondrial	—NA—
ahe-miR034	miR172	AGAAUCUUGAUGAUGCUGCAG	21	DN8633_c1_g1_i1	DN56487_c0_g1_i1	1	—NA—	—NA—
ahe-miR035	miR172	AGAAUCUUGAUGAUGCUGC	19	DN8633_c1_g1_i1	DN56487_c0_g1_i1	1	—NA—	—NA—
ahe-miR036	miR172	GAAUCUUGAUGAUGCUGCAU	20	DN8633_c1_g1_i1	DN56487_c0_g1_i1	0	—NA—	—NA—
ahe-miR037	miR172	AUCUUGAUGAUGCUGCAUCGGC	22	DN8633_c1_g1_i1	DN56487_c0_g1_i1	0.5	—NA—	—NA—
ahe-miR038	miR319	UUGAGGUGAUUAAUAAAUAUUU	22	DN34963_c0_g1_i1	DN8925_c0_g1_i4	2	43 kDa postsynaptic protein	DN67158_c0_g1_i1
ahe-miR039	miR393	UUGGGAUCAUGCUAUCCCUUUG	22	DN10344_c0_g1_i2	DN17619_c0_g1_i1	3	—NA—	—NA—
ahe-miR040	miR393	AAAGGGAUCGCAUUGAUCCCAA	22	DN10344_c0_g1_i2	DN1938_c1_g1_i2	2	protein AUXIN SIGNALING F-BOX 2	—NA—
ahe-miR041	miR394	AGGUGGGCAUACUGCCAACUGA	22	DN2457_c1_g1_i3	DN21191_c0_g1_i4	3	uncharacterized protein LOC21394998	—NA—
ahe-miR042	miR396	UUCCACAGCUUUCUUGAACUG	21	DN24771_c0_g1_i1	DN17464_c0_g1_i5	2	—NA—	—NA—
ahe-miR043	miR396	UUCCACAGCUUUCUUGAACU	20	DN24771_c0_g1_i1	DN17464_c0_g1_i5	2	—NA—	—NA—
ahe-miR044	miR396	UUCCACAGCUUUCUUGAACUU	21	DN20110_c0_g1_i1	DN17464_c0_g1_i5	2	—NA—	—NA—
ahe-miR045	miR396	CUCAAGAAAGCUGUGGGAGA	20	DN20110_c0_g1_i1	DN21027_c0_g1_i1	2	—NA—	—NA—
ahe-miR046	miR396	GCUCAAGAAAGCUGUGGGAGA	21	DN20110_c0_g1_i1	DN21027_c0_g1_i1	2	—NA—	—NA—
ahe-miR047	miR396	UCCACAGCUUUCUUGAACUG	20	DN24771_c0_g1_i1	DN37359_c0_g2_i1	2	alanine-glyoxylate aminotransferase 1	—NA—

(*continued on next page*)

Table 3 (*continued*)

Predicted_miR_ID	miRNA family	Predicted miRNA seq_RNA form	Nucleotide length	Precursor miRNA Transcript_ID	Target transcript ID	Expectation value	Description	eTMs
ahe-miR048	miR828	UCAAAUGAGCAUCUCAGAAAAC	22	DN18474_c0_g2_i1	DN4668_c0_g1_i5	2	tRNA(adenine(34)) deaminase, chloroplastic	—NA—
ahe-miR049	miR2916	UGAGUCAAAUUAAGCCGCAGGC	22	DN21362_c1_g1_i2	DN1485_c0_g1_i10	0	hypothetical protein RclHR1_00930026	—NA—
ahe-miR050	miRf10081-akr	GAAUCUUGAUGAUGCUGCAUCGGC	24	DN8633_c1_g1_i1	DN5647_c0_g1_i1	0	—NA—	—NA—
ahe-miR051	miRf10509-akr	GAUUGAGCCGCGCCAAUAUCACUU	24	DN66416_c0_g1_i1	DN4118_c0_g1_i1	0	scarecrow-like protein 22	—NA—
ahe-miR052	miRf10553-akr	UGGGAAUCUUGAUGAUGCUG	20	DN33154_c0_g1_i2	DN5647_c0_g1_i13	0.5	floral homeotic protein APETALA 2 isoform X1	—NA—
ahe-miR053	miRf10564-akr	GAGAAUCUUGAUGAUGCUGC	20	DN8633_c1_g1_i1	DN38291_c0_g1_i1	1	pentatricopeptide repeat-containing protein At1g20300, mitochondrial	—NA—

significantly lower expression in leaf primordia than their target genes. On the contrary, miRNAs expressed higher than their corresponding target genes in the matured leaves. The remaining four miRNAs: ahe-miR001, ahe miR028, ahe-miR034, and ahe-miR042, did not express in the leaves, while its target genes DN5506, DN38291, DN56487, DN17464 respectively showed significant expression.

4. Discussion

Jack (*A. heterophyllus*) is an underutilized wild fruit tree species. It has gained significant popularity in recent years, mainly as an alternative to meat [40]. As a result, Jack, which grew naturally along the roadsides or in woodlots, is gradually becoming an orchard crop [2]. Jack fruit meets various food and non-food purposes with varying end-users preferences. However, genetic improvement in Jack aimed at catering to different end-users' needs and a worldwide concern over the ever-increasing loss of its diversity needs more attention from the scientific community [41].

The next-generation sequencing (NGS) technologies provide novel opportunities to develop enormous genomic resources prerequisite for efficient breeding, understanding the genetic basis of trait variation, and spearheading diversity conservation strategies in plants and trees [42]. Towards this, we sequenced a normalized cDNA paired-end sequencing library, prepared from an equimolar mixture of total RNA extracted from multiple tissues of *A. heterophyllus*, using an Illumina NextSeq1000 platform. A normalized cDNA library prepared from various tissues enables adequate sampling of transcript complexity besides maximizing the probability of detecting less abundant mRNA [43]. The paired-end sequencing of the cDNA library yields reads with significant overlaps critical for their assembly, particularly in non-model tree species lacking reference genome sequence [44]. The NGS generated 42,928,887 paired-end raw reads, out of which 41,549,555 (96.8%) were retained after quality filtering, indicating that the library was of sufficient quality for precise sequencing. The high-quality reads were assembled into 80,411 unigenes using the Trinity assembly strategy, optimally suitable for transcriptome construction without a reference genome [45]. A high N50 value of 1265 bp and an average length of 682 bp confirmed the high quality of assembly data. The assembly consistency assessed through Bowtie2 and BWA alignment tools indicated 97.08 and 94.27% reads mapping back to the assembled transcriptome, which is much higher than retrieved in many plant and tree species [46,47]. Quality assessment of the transcriptome assembly using the eukaryota and viridiplantae databases of BUSCO genes indicated that the assembled dataset comprised 88.2% of BUSCO genes in eukaryota and 77.4% of BUSCO genes in the viridiplantae dataset. Besides, 7.5% and 16.5% of genes were fragmented, and only 4.3% and 6.1% were missing from the eukaryota and viridiplantae datasets. BUSCO recovery tends to be highest when the whole organism is used to generate the assembly, compared to those assembled from a selected number of tissues [48]. Nevertheless, the missing of only a small proportion of genes from the eukaryota and viridiplantae datasets indicated that the assembled dataset achieved a reasonable degree of completeness and was comparable to many of the recent studies [49].

CDS prediction using TransDecoder indicated that approximately 80% of the unigenes coded for medium to long ORFs. Among the predicted CDSs, 95.5% had at least one significant match in the nr database, indicating that most CDSs code for proteins. The CDSs that had no significant matches may lack a known conserved functional domain or represent non-coding RNAs, or maybe very short [50]. The unavailability of a sufficient number of *A. heterophyllus* lineage specific genes in the databases may also be the possible reason for the failure to get a significant match. The study revealed that *Morus notabilis* genes had the maximum similarity (71.8%) with *A. heterophyllus* CDSs, indicating that the genome sequence of *Morus notabilis* may serve as a reference in the future transcriptomic study in *A. heterophyllus*. Gene ontology analysis classified 53.3% of *A. heterophyllus* CDSs majorly into cellular

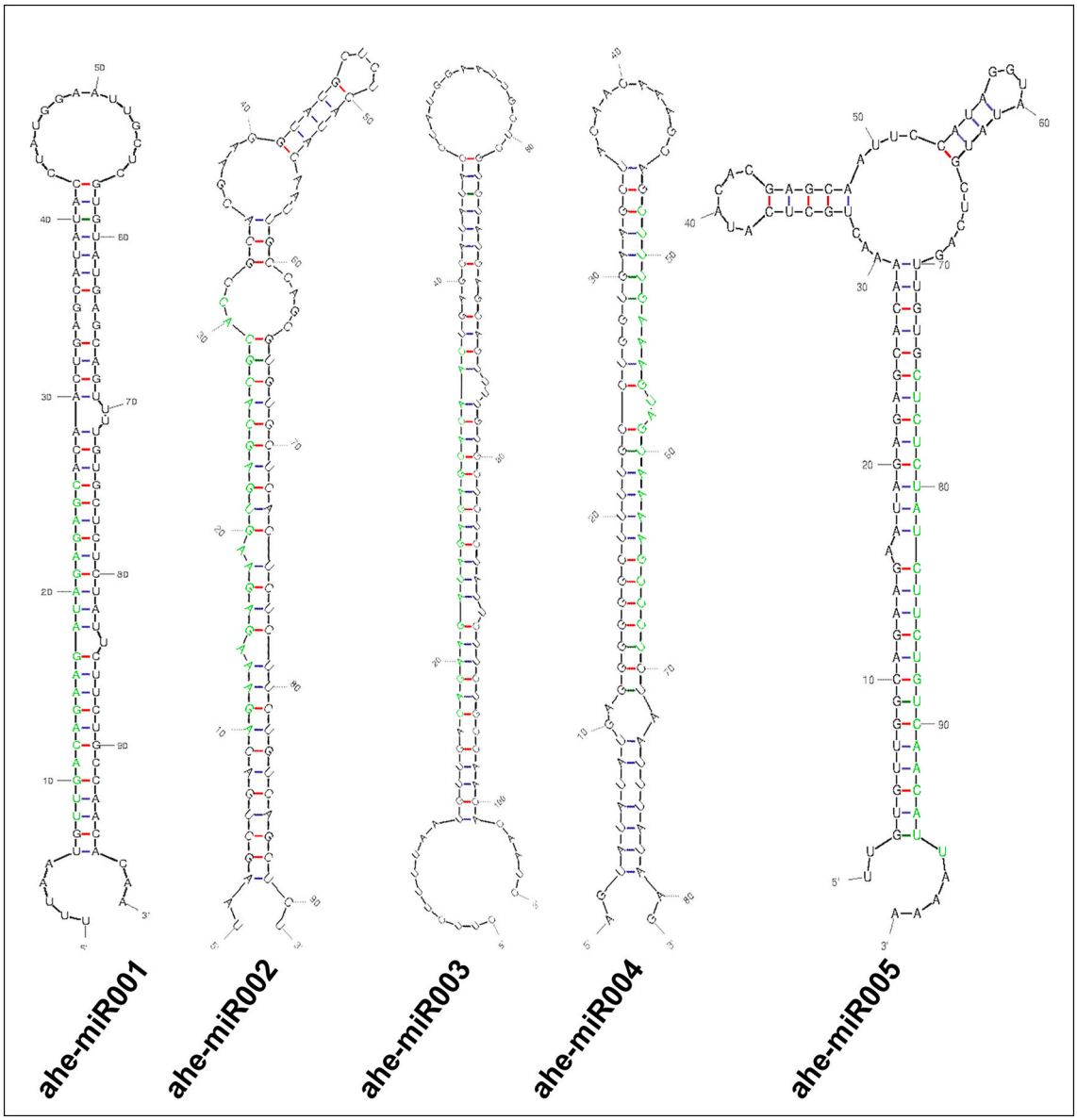

Fig. 7. Secondary structures of selected pre-miRNA sequences in *Artocarpus heterophyllus.*

component, molecular function, and biological process categories, among which the intrinsic compound of membrane protein, heterocyclic, organic cyclic and ion binding protein, and organic substance metabolic process, primary metabolic process, cellular metabolic process, and nitrogen compound metabolic process accounted for the majority of CDSs. The above findings are consistent with various other transcriptome studies in plants [50,51].

We identified 16,853 perfect SSRs from the assembled unigenes, accounting for one SSR per 6.4 kbp unigene sequence, which is significantly higher than in several other perennial tree species like *Coffea canephora* (7.73 kbp) [52] but lower than in *Liquidambar formosana* (5.28 kbp) [53], and *Hevea brasiliensis* (0.28 kbp) [54]. Genome-size, redundancy in the unigenes, and tools and parameters employed to detect SSRs often lead to these variations [55]. A large proportion of the genic-SSRs identified in the study comprised repeats of trinucleotides (80.57%), similar to the trend observed in several other plant species [41,56,57]. Continuous expansion or contraction of dinucleotide repeats to suppress the deleterious effects of frame-shift mutations possibly explain trinucleotides' abundance in the coding regions [58]. The number of reiterations of the repeat motifs varied from 4 to 22, corroborating the earlier report in *Morus alba* [59] of the Artocarpus lineage.

Transcriptional regulation of gene expression is primarily affected by TFs and TRs. Identification of novel TFs and TRs provide new insights into context-dependent gene expression and generation of novel phenotype [60]. We cataloged 2741 unigenes into 69 TF families and 489 unigenes into 29 TR families using PlantTFcat online tool. Like other studies, C2H2-type zinc finger proteins were the most abundant TFs in *A. heterophyllus*. These genes have well-proven roles in plant growth and development. Moreover, they are considered the master regulators of abiotic stress responses in plants [61]. Among the TRs, the plant homeodomain (PHD) was most abundant. Plant homeodomain finger proteins are extensively present in plants and play crucial roles in chromatin remodelling and transcriptional regulation. They play diverse roles in plant growth and development and are considered an important source of candidate genes for genetic engineering-mediated trait manipulation in plants [62].

A majority of the mature miRNAs in plant species are evolutionarily conserved and usually, target conserved homologous genes in diverse plant species [63]. Taking advantage of this fact, we identified 53

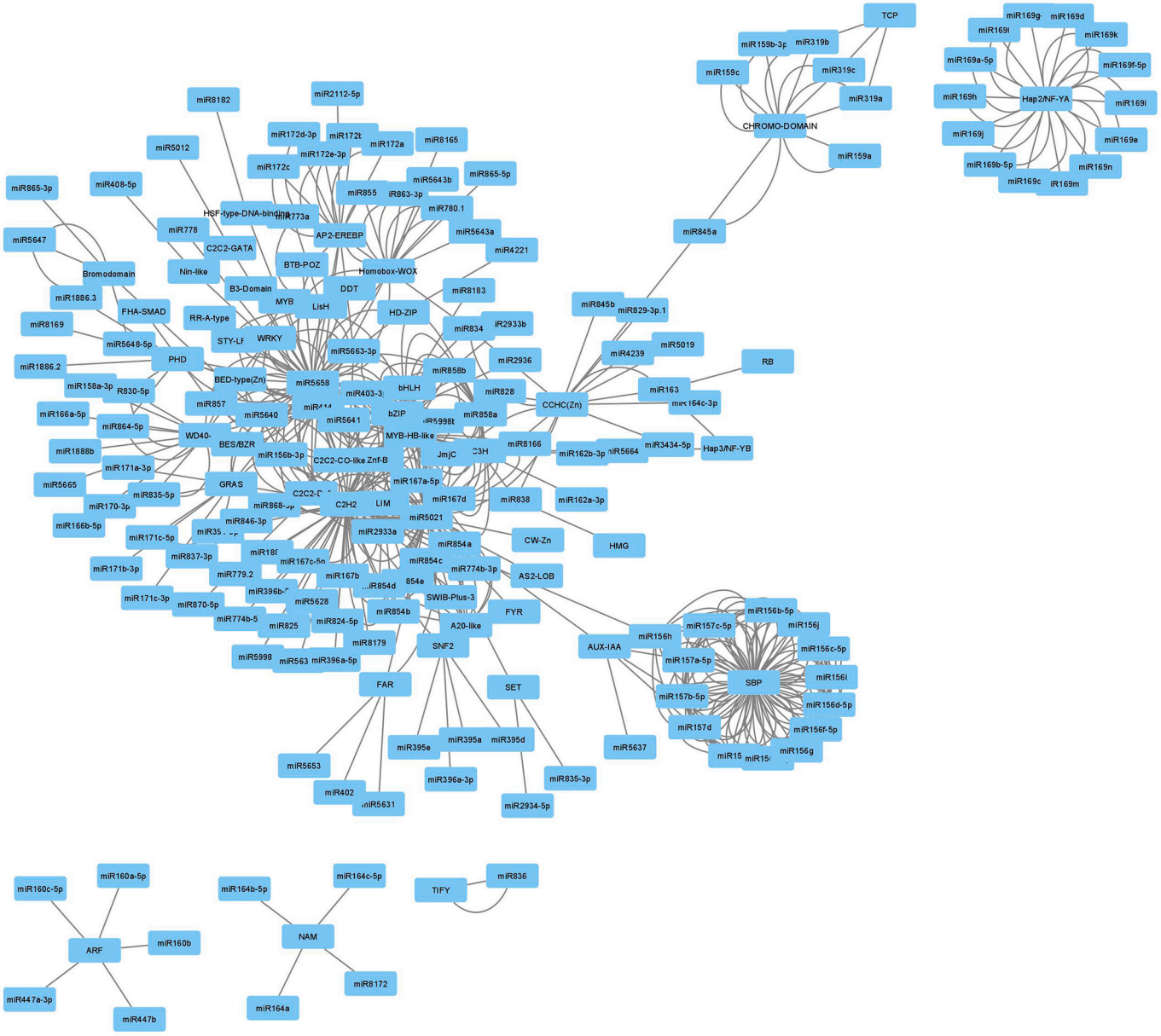

Fig. 8. Cystoscope network showing the interaction between *Artocarpus heterophyllus* transcription factors and miRNAs.

miRNAs belonging to 19 conserved miRNA families following an optimal computational identification strategy. These miRNAs target 31different unigenes, mostly coding for TFs involved in various developmental pathways. Moreover, we also predicted several novel functions for miRNAs in *A. heterophyllus,* which reflects that at least some conserved miRNAs are regulating new targets in addition to the well-documented conserved targets. While analyzing our data, we identified various TFs as miRNA-targets like squamosa promoter-binding –like proteins, auxin response factors, homeobox-leucine zipper proteins, ethylene-responsive transcription factors, etc. Analysis of the gene-TFs-miRNAs regulatory network allowed us to draw a comprehensive picture of the involvement of TFs and miRNAs in the regulation of diverse processes leading to plant growth and development and biotic and abiotic responses.

In recent years, discovering lncRNAs and elucidating lncRNA-target interactions have become important research activities in plants. However, the lack of widespread sequence-level conservation and limited knowledge about the lncRNA-target interaction interfaces make the task very challenging [64]. Therefore, steady progress in characterizing lncRNAs and their targets from different species is critical to their efficient identification. Through standard *in-silico* computational approaches, we identified 5350 potential lncRNAs targeting a similar number of mRNAs in *A. heterophyllus*. Moreover, we also placed three eTMs in the lncRNA-miRNA pairs using the lncRNAs and miRNAs identified in the study. A detailed analysis of these genetic elements would help develop novel functional networks and modulation for end-use-based improvement in *A. heterophyllus*.

The analysis of the expression patterns of some selected miRNAs identified in the study *vis-à-vis* their target genes helped us experimentally validate and provide insight into the functional cues of these miRNAs. Of the seven miRNAs analyzed, three miRNAs targeting auxin response factor, homeobox leucine zipper, and an unknown gene showed an inverse relationship with the expression patterns of their target genes. These miRNAs behaved differently at different stages of leaf development. At the primordial stage, the expression of all three miRNAs was substantially lower than at the maturity stage. Several studies have evidenced a higher built-up of auxin response factors and homeobox leucine zipper proteins during the early stage of leaf

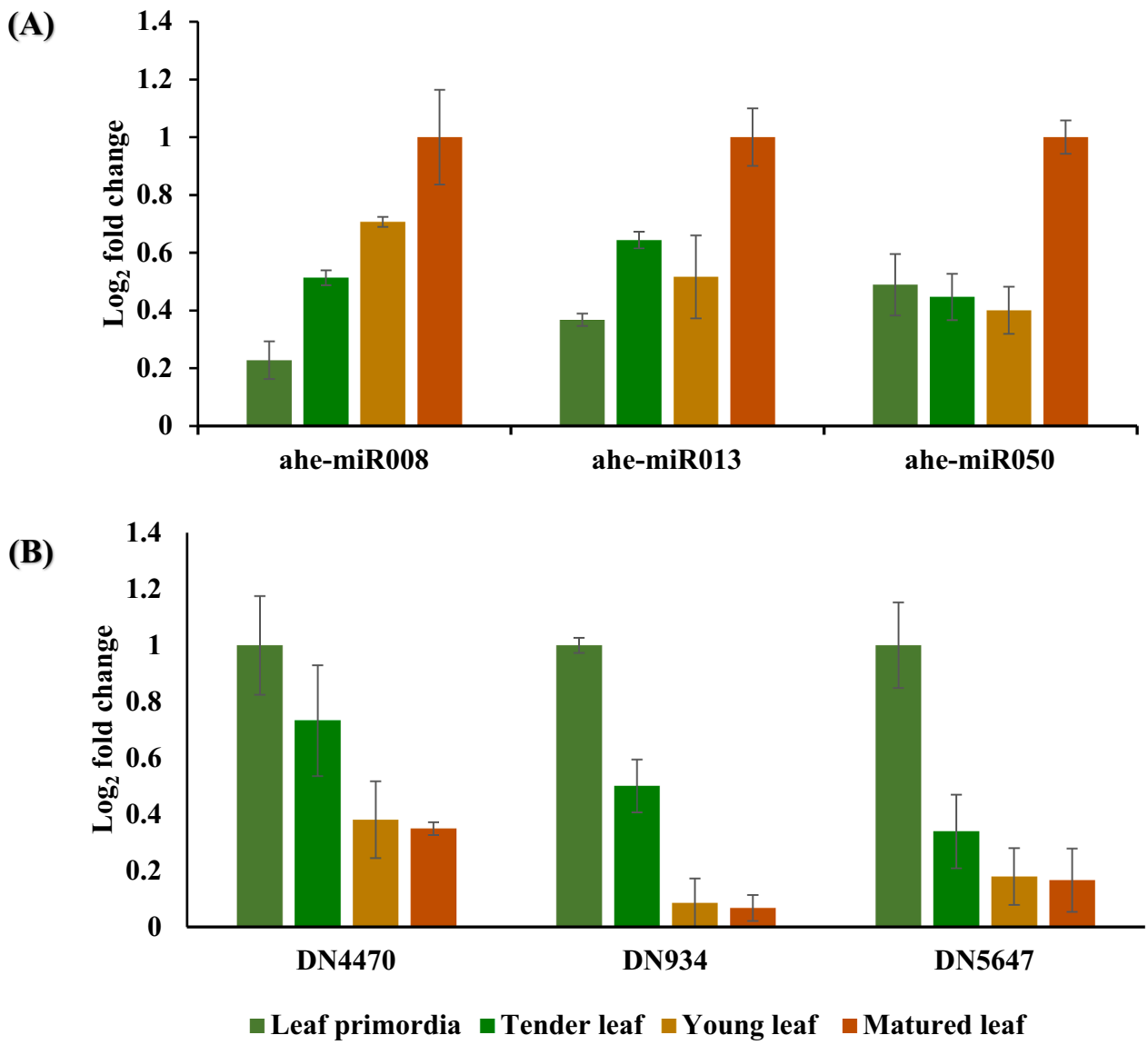

Fig. 9. Quantitative real-time PCR expression analysis, (A) miRNAs (ahe-miR008, ahe-miR013, and ahe-miR050), and (B) target genes (DN4470, DN934 and DN5647) at different stages of leaf development in *Artocarpus heterophyllus*.

development than at the maturity stage [65]. The study indicates that the miRNAs predicted in *A. heterophyllus* are valid, and detailed functional studies of these riboregulators would help better understand various cellular and metabolic functions.

5. Conclusion

Through global transcriptome sequencing and analysis in *A. heterophyllus*, we identified 80,411 unigenes from which we predicted 64,215 CDSs and discovered 16,853 perfect SSRs. In addition, we identified 2741 TFs, 489 TRs, 53 miRNAs, 25,953 potential lncRNAs and placed 03 functional eTMs in different lncRNA-miRNA pairs. Through network analysis involving genes, TFs, and miRNAs, we deduced a comprehensive picture of TFs and miRNAs-mediated regulation in *A. heterophyllus*. Finally, we validated a set of *in-silico* identified miRNAs by comparing their expression patterns *vis-à-vis* their corresponding target genes at different leaf developmental stages. The genomic resources developed in the study would boost end-use-based improvement and conservation studies for sustainable use of the species.

Funding

This study was conducted under the project "Development of transcriptome-based resources for indigenous agri-horticultural crops of eastern India," funded by ICAR - Indian Institute of Agricultural Biotechnology, Ranchi 834 003, Jharkhand, India.

Authors' contributions

BKS conceived the project and wrote the manuscript. BKS and KUT designed the experiments. BP, DKS, KUT performed the experiments. SKB, AP, SK provided valuable inputs in the research work. KUT, JB, DCM, AD analyzed the data. TRS and AP coordinated the project. All the authors read and approved the manuscript.

Declaration of Competing Interest

The authors of this manuscript have no competing interests.

Acknowledgments

We sincerely acknowledge Director, ICAR - Indian Institute of Agricultural Biotechnology, Ranchi 834 003, Jharkhand, India, for providing financial support and the facilities to carry out this research work. We also gratefully acknowledge the Director, ICAR – National Bureau of Plant Genetic Resources, New Delhi 110 012, India, to provide plant materials of *A. heterophyllus* for the study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110356.

References

[1] K. Kishore, Phenological growth stages of jackfruit (*Artocarpus heterophyllus*) according to the extended BBCH scale, Ann. Appl. Biol. 172 (2018) 366–374, https://doi.org/10.1111/aab.12427.
[2] S.L. Jagadeesh, B.S. Reddy, N. Basavaraj, G.S.K. Swamy, K. Gorbal, L. Hegde, G.S. V. Raghavan, S.T. Kajjidoni, Inter tree variability for fruit quality in jackfruit selections of Western Ghats of India, Sci. Hortic. 112 (2007) 382–387, https://doi.org/10.1016/j.scienta.2007.01.016.
[3] C.R. Elevitch, H.I. Manner, Artocarpus heterophyllus (jackfruit), ver. 1.1v, in: C. R. Elevitch (Ed.), Species Profiles for Pacific Island Agroforestry. Permanent Agriculture Resources (PAR), Holualoa, Hawaii, 2006.
[4] R.A.S.N. Ranasinghe, S.D.T. Maduwanthi, R.A.U.J. Marapana, Nutritional and health benefits of jackfruit (*Artocarpus heterophyllus* Lam.): a review, Int. J. Food Sci. (2019), https://doi.org/10.1155/2019/4327183.
[5] L. Feijoo-Siota, T.G. Villa, Native and biotechnologically engineered plant proteases with industrial applications, Food Bioprocess Technol. 4 (2011) 1066–1088, https://doi.org/10.1007/s11947-010-0431-4.
[6] M.T. Hossain, M.M. Hossain, M. Sarker, A.N. Shuvo, M.M. Alam, M.S. Rahman, Development and quality evaluation of bread supplemented with jackfruit seed flour, IJNFS 3 (2014) 484–487, https://doi.org/10.11648/j.ijnfs.20140305.28.
[7] M.S. Madruga, F.S.M. de Albuquerque, I.R.A. Silva, D.S. Do Amaral, M. Magnani, V. Q. Neto, Chemical, morphological and functional properties of Brazilian jackfruit (*Artocarpus heterophyllus* L.) seeds starch, Food Chem. 15 (2014) 440–445, https://doi.org/10.1016/j.foodchem.2013.08.003.
[8] F.P. Spada, P.P.M. da Silva, G.F. Mandro, G.B. Margiotta, M.H.F. Spoto, S. G. Canniatti-Brazaca, Physicochemical characteristics and high sensory acceptability in cappuccinos made with jackfruit seeds replacing cocoa powder, PLoS One 13 (2018), e0197654, https://doi.org/10.1371/journal.pone.0197654.
[9] J. Ahmed, L. Thomas, Oscillating rheology of jackfruit (*Artocarpus heterophyllus*) seed flour dough in relation to different particle size, J. Food Process Eng. 43 (2020), e13558, https://doi.org/10.1111/jfpe.13558.
[10] Y. Zhang, B. Li, F. Xu, S. He, Y. Zhang, L. Sun, K. Zhu, S. Li, G. Wu, L. Tan, Jackfruit starch: composition, structure, functional properties, modifications and applications, Trends Food Sci. Technol. 107 (2021) 268–283, https://doi.org/10.1016/j.tifs.2020.10.041.
[11] S.B. Swami, N.J. Thakor, P.M. Haldankar, S.B. Kalse, Jackfruit and its many functional components as related to human health: a review, Compr. Rev. Food Sci. Food Saf. 11 (2012) 565–576, https://doi.org/10.1111/j.1541-4337.2012.00210.x.
[12] A.K. Gupta, M.A. Rather, A.K. Jha, A. Shashank, S. Singhal, M. Sharma, U. Pathak, D. Sharma, A. Mastinu, *Artocarpus lakoocha* roxb. And *Artocarpus heterophyllus* lam. flowers: new sources of bioactive compounds, Plants 9 (2020) 1329, https://doi.org/10.3390/plants9101329.
[13] M. Gundogdu, K. Ozrenk, S. Ercisli, T. Kan, O. Kodad, A. Hegedus, Organic acids, sugars, vitamin C content and some pomological characteristics of eleven hawthorn species (*Crataegus* spp.) from Turkey, Biol. Res. 47 (2014) 21, https://doi.org/10.1186/0717-6287-47-21.
[14] S.P. Engin, C. Mert, The effects of harvesting time on the physicochemical components of aronia berry, Turk. J. Agric. For. 44 (2020) 361–370, https://doi.org/10.3906/tar-1903-130.
[15] V. Kaskoniene, K. Bimbiraite-Surviliene, P. Kaskonas, N. Tiso, L. Cesoniene, R. Daubaras, A.S. Maruska, Changes in the biochemical compounds of *Vaccinium myrtillus*, *Vaccinium vitisidaea*, and forest litter collected from various forest types, Turk. J. Agric. For. 44 (2020) 557–566, https://doi.org/10.3906/tar-1912-41.
[16] R.J. Campbell, N. Ledesma, The Exotic jackfruit: Growing the world's Largest Fruit, Fairchild Tropical Garden, Coral Gables, FL, 2003.
[17] V.A. Bapat, U.B. Jagtap, S.B. Ghag, T.R. Ganapathi, Molecular approaches for the improvement of under-researched tropical fruit trees: jackfruit, guava, and custard apple, Int. J. Fruit Sci. 20 (2020) 233–281, https://doi.org/10.1080/15538362.2019.1621236.
[18] S. Moeinizade, Y. Han, H. Pham, G. Hu, L. Wang, A look-ahead Monte Carlo simulation method for improving parental selection in trait introgression, Sci. Rep. 3918 (2021), https://doi.org/10.1038/s41598-021-83634-x.
[19] A.T. Djami-Tchatchou, N. Sanan-Mishra, K. Ntushelo, I.A. Dubery, Functional roles of microRNAs in agronomically important plants—potential as targets for crop improvement and protection, Front. Plant Sci. (2017), https://doi.org/10.3389/fpls.2017.00378.
[20] J. Tang, C. Chu, MicroRNAs in crop improvement: fine-tuners for complex traits, Nat. Plants 17077 (2017), https://doi.org/10.1038/nplants.2017.77.
[21] B. Zhang, Q. Wang, MicroRNA-based biotechnology for plant improvement, J. Cell. Physiol. 230 (2015) 1–15, https://doi.org/10.1002/jcp.24685.
[22] S.K. Sahu, M. Liu, A. Yssel, R. Kariba, S. Muthemba, S. Jiang, B. Song, P.S. Hendre, A. Muchugi, R. Jamnadass, S.M. Kao, J. Featherston, N.J.C. Zerega, X. Xu, H. Yang, A.V. Deynze, Y.V. Peer, X. Liu, H. Liu, Draft genomes of two Artocarpus Plants, Jackfruit (*A. heterophyllus*) and Breadfruit (*A. altilis*), Genes (Basel) 11 (2019) 27, https://doi.org/10.3390/genes11010027.
[23] M. Mathiazhagan, B. Chidambara, L.R. Hunashikatti, K.V. Ravishankar, Genomic approaches for improvement of tropical fruits: fruit quality, shelf life and nutrient content, Genes (Basel) 12 (2021) 1881, https://doi.org/10.3390/genes12121881.
[24] A. Kumar, T. Anju, S. Kumar, S.S. Chhapekar, S. Sreedharan, S. Singh, S.R. Choi, N. Ramchiary, Y.P. Lim, Integrating omics and gene editing tools for rapid improvement of traditional food plants for diversified and sustainable food security, Int. J. Mol. Sci. 22 (2021) 8093, https://doi.org/10.3390/ijms22158093.
[25] X. Meng, J. Xu, M. Zhang, R. Du, W. Zhao, Q. Zeng, Z. Tu, J. Chen, B. Chen, Third-generation sequencing and metabolome analysis reveal candidate genes and metabolites with altered levels in albino jackfruit seedlings, BMC Genomics 22 (2021) 543, https://doi.org/10.1186/s12864-021-07873-y.
[26] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120, https://doi.org/10.1093/bioinformatics/btu170.
[27] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. de Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652, https://doi.org/10.1038/nbt.1883.
[28] W. Li, A. Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics 22 (2006) 1658–1659, https://doi.org/10.1093/bioinformatics/btl158.
[29] H. Li, R. Durbin, Fast and accurate short read alignment with burrows-wheeler transform, Bioinformatics 25 (2009) 754–1760, https://doi.org/10.1093/bioinformatics/btp324.
[30] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) (2009) 2078–2079, https://doi.org/10.1093/bioinformatics/btp352.
[31] S. Conesa, J.M. Götz, J. García-Gómez, M. Terol, M. Talón, Robles, Blast2GO: a universal tool for annotation, visualization, and analysis in functional genomics research, Bioinformatics 21 (2005) 3674–3676, https://doi.org/10.1093/bioinformatics/bti610.
[32] T. Thiel, W. Michalek, R.K. Varshney, A. Graner, Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.), Theor. Appl. Genet. 106 (2003) 411–422, https://doi.org/10.1007/s00122-002-1031-0.
[33] K. Okonechnikov, O. Golosova, M. Fursov, UGENE team, Unipro UGENE: a unified bioinformatics toolkit, Bioinformatics 28 (2012) 1166–1167, https://doi.org/10.1093/bioinformatics/bts091.
[34] M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction, Nucleic Acids Res. 31 (2003) 3406–3415, https://doi.org/10.1093/nar/gkg595.
[35] B.C. Meyers, M.J. Axtell, B. Bartel, D.P. Bartel, D. Baulcombe, J.L. Bowman, X. Cao, J.C. Carrington, X. Chen, P.J. Green, S. Griffiths-Jones, S.E. Jacobsen, A.C. Mallory, R.A. Martienssen, R.S. Poethig, Y. Qi, H. Vaucheret, O. Voinnet, Y. Watanabe, D. Weigel, J.K. Zhu, Criteria for annotation of plant MicroRNAs, Plant Cell 20 (2008) 3186–3190, https://doi.org/10.1105/tpc.108.064311.
[36] M.J. Axtell, B.C. Meyers, Revisiting criteria for plant microRNA annotation in the era of big data, Plant Cell 30 (2018) 272–284, https://doi.org/10.1105/tpc.17.00851.
[37] X. Dai, Z. Zhuang, P.X. Zhao, psRNATarget: a plant small RNA target analysis server (2017 release), Nucleic Acids Res. 46 (2018) W49–W54, https://doi.org/10.1093/nar/gky316.
[38] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.
[39] S.A. Bustin, V. Benes, J.A. Garson, J. Hellemans, J. Huggett, M. Kubista, R. Mueller, T. Nolan, M.W. Pfaffl, G.L. Shipley, J. Vandesompele, C.T. Wittwer, The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments, Clin. Chem. 55 (2009) 611–622, https://doi.org/10.1373/clinchem.2008.112797.
[40] H.L. Anila, S. Divakar, Functional properties of raw jackfruit based textured vegetable protein (TVP), Int. J. Food Sci. Nutr. 3 (2018) 52–54.
[41] D.K. Singh, A. Pandey, S.B. Choudhary, S. Kumar, K.U. Tribhuvan, D.C. Mishra, J. Bhati, M. Kumar, J.B. Tomar, S.K. Bishnoi, M.A. Mallick, V.P. Bhadana, T. R. Sharma, A. Pattanayak, B.K. Singh, Development of genic-SSR markers and their application in revealing genetic diversity and population structure in an Eastern and North-Eastern Indian collection of Jack (*Artocarpus heterophyllus* Lam.), Ecol. Indic. 131 (2021) 111–120, https://doi.org/10.1016/j.indcrop.2018.01.023.
[42] P. Yadav, E. Vaidya, R. Rani, N.K. Yadav, B.K. Singh, P.K. Rai, D. Singh, Recent perspective of next-generation sequencing: applications in molecular plant biology and crop improvement, Proc. Natl. Acad. Sci., India - Sect. B: Biol. Sci. 88 (2) (2016) 435–449, https://doi.org/10.1007/s40011-016-0770-7.

[43] N.V. Hoang, A. Furtado, V. Perlo, F.C. Botha, R.J. Henry, The impact of cDNA normalization on long-read sequencing of a complex transcriptome, Front. Genet. (2019), https://doi.org/10.3389/fgene.2019.00654.
[44] J.A. Martin, Z. Wang, Next-generation transcriptome assembly, Nat. Rev. Genet. 12 (10) (2011) 671–682, https://doi.org/10.1038/nrg3068.
[45] F. Amil-Ruiz, A.M. Herruzo-Ruiz, C. Fuentes-Almagro, C. Baena-Angulo, J. M. Jiménez-Pastor, J. Blasco, J. Alhama, C. Michán, Constructing a *de novo* transcriptome and a reference proteome for the bivalve *Scrobicularia plana*: comparative analysis of different assembly strategies and proteomic analysis, Genomics 113 (2021) 1543–1553, https://doi.org/10.1016/j.ygeno.2021.03.025.
[46] U.J. Siregar, A. Nugroho, H. Shabrina, F. Indriani, A. Damayanti, D.D. Matra, De novo transcriptome assembly data for sengon (*Falcataria moluccana*) trees displaying resistance and susceptibility to *boktor* stem borers (*Xystrocera festiva* Pascoe), BMC Res. Notes 14 (2021) 261, https://doi.org/10.1186/s13104-021-05675-9.
[47] H.Z. Li, X. Gao, X.Y. Li, Q.J. Chen, J. Dong, W.C. Zhao, Evaluation of assembly strategies using RNA-seq data associated with grain development of wheat (*Triticum aestivum* L.), PLoS One 8 (12) (2013), e83530, https://doi.org/10.1371/journal.pone.0083530.
[48] M. Carruthers, A.A. Yurchenko, J.J. Augley, C.E. Adams, P. Herzyk, K.R. Elmer, De novo transcriptome assembly, annotation, and comparison of four ecological and evolutionary model salmonid fish species, BMC Genomics 19 (2018) 32, https://doi.org/10.1186/s12864-017-4379-x.
[49] E.J. Carpenter, N. Matasci, S. Ayyampalayam, S. Wu, J. Sun, J. Yu, F.R. Jimenez Vieira, C. Bowler, R.G. Dorrell, M.A. Gitzendanner, L. Li, W. Du, K.K. Ullrich, N. J. Wickett, T.J. Barkmann, M.S. Barker, J.H. Leebens-Mack, G. Ka-Shu Wong, Access to RNA-sequencing data from 1,173 plant species: the 1000 plant transcriptomes initiative (1KP), Gigascience 8 (2019) 1–7, https://doi.org/10.1093/gigascience/giz126.
[50] G. Wu, L. Zhang, Y. Yin, J. Wu, L. Yu, Y. Zhou, M. Li, Sequencing, de novo assembly and comparative analysis of *Raphanus sativus* transcriptome, Front. Plant Sci. (2015), https://doi.org/10.3389/fpls.2015.00198.
[51] A.B. Mazumdar, S. Chattopadhyay, Sequencing, de novo assembly, functional annotation and analysis of *Phyllanthus amarus* leaf transcriptome using the Illumina platform, Front. Plant Sci. 28 (6) (2016) 1199, https://doi.org/10.3389/fpls.2015.01199.
[52] V. Poncet, M. Rondeau, C. Tranchant, A. Cayrel, S. Hamon, A. Kochko, P. Hamon, SSR mining in coffee tree EST databases: potential use of EST-SSRs as markers for the *Coffea* genus, Mol. Gen. Genomics. 276 (2006) 436–449, https://doi.org/10.1007/s00438-006-0153-5.
[53] R. Sun, F. Lin, P. Huang, Y. Zheng, Moderate genetic diversity and genetic differentiation in the relict tree *Liquidambar formosana* Hance revealed by genic simple sequence repeat markers, Front. Plant Sci. 7 (2016) 1411, https://doi.org/10.3389/fpls.2016.01411.
[54] D. Li, Z. Deng, B. Qin, X. Liu, Z. Men, De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.), BMC Genomics 13 (2012) 192, https://doi.org/10.1186/1471-2164-13-192.
[55] B.K. Singh, D.C. Mishra, S. Yadav, S. Ambawat, E. Vaidya, K.U. Tribhuvan, A. Kumar, S. Kumar, S. Kumar, K.K. Chaturvedi, R. Rani, P. Yadav, A. Rai, P.K. Rai, V.V. Singh, D. Singh, Identification, characterization, validation and cross-species amplification of genic-SSRs in Indian mustard (*Brassica juncea*), J. Plant Biochem. Biotechnol. 25 (4) (2016) 410–420, https://doi.org/10.1007/s13562-016-0353-y.
[56] Y.K. Ahn, S. Tripathi, Y.I. Cho, J.H. Kim, H.E. Lee, D.S. Kim, J.G. Woo, M.C. Cho, De novo transcriptome assembly and novel microsatellite marker information in *Capsicum annuum* varieties Saengryeg 211 and Saengryeg 213, Bot. Stud. 54 (2013) 58. http://www.as-botanicalstudies.com/content/54/1/58.
[57] B.K. Singh, S.B. Choudhary, S. Yadav, E.V. Malhotra, R. Rani, S. Ambawat, A. Priyamedha, R. Pandey, S. Kumar, H.K. Kumar, D.K. Sharma, P.K. Rai Singh, Genetic structure identification and assessment of interrelationships between Brassica and allied genera using newly developed genic-SSRs of Indian Mustard (***Brassica juncea*** L.), Ind. Crop. Prod. 113 (2018) 111–120, https://doi.org/10.1016/j.indcrop.2018.01.023.
[58] D. Xin, J. Sun, J. Wang, H. Jiang, G. Hu, C. Liu, Q. Chen, Identification and characterization of SSRs from soybean (*Glycine max*) ESTs, Mol. Biol. Rep. 39 (2012) 9047–9057, https://doi.org/10.1007/s11033-012-1776-8.
[59] M. Thumilan, R.S. Sajeevan, J. Biradar, T. Madhuri, K.N. Nataraja, S.M. Sreeman, S.K. Parida, Development and characterization of genic SSR markers from Indian Mulberry transcriptome and their transferability to related species of Moraceae, PLoS One 11 (9) (2016), e0162909, https://doi.org/10.1371/journal.pone.0162909.
[60] F.J. Buitrago-Flórez, S. Restrepo, D.M. Riaño-Pachón, Identification of transcription factor genes and their correlation with the high diversity of Stramenopiles, PLoS One 9 (11) (2014), e111841, https://doi.org/10.1371/journal.pone.0111841.
[61] H. Guoliang, L. Chaoxia, G. Jianrong, Q. Ziqi, S. Na, Q. Nianwei, W. Baoshan, C2H2 zinc finger proteins: master regulators of abiotic stress responses in plants, Front. Plant Sci. 11 (2020) 115, https://doi.org/10.3389/fpls.2020.00115.
[62] C.C. Alam, H.L. Liu, K. Ge, Y.Q. Batool, Y.H. Yang, Lu, genome-wide survey, evolution, and expression analysis of PHD finger genes reveal their diverse roles during the development and abiotic stress responses in *Brassica rapa* L, BMC Genomics 20 (2019) 773, https://doi.org/10.1186/s12864-019-6080-8.
[63] B. Zhang, X. Pan, C.H. Cannon, G.P. Cobb, T.A. Anderson, Conservation and divergence of plant microRNA genes, Plant J. 46 (2) (2006) 243–259, https://doi.org/10.1111/j.1365-313X.2006.02697.x.
[64] B. Hikmet, K.S. Biyiklioglu, C.S. Busra, Long non-coding RNA in plants in the era of reference sequences, Front. Plant Sci. 11 (2020) 276, https://doi.org/10.3389/fpls.2020.00276.
[65] M.F. Perotti, P.A. Ribone, R.L. Chan, Plant transcription factors from the homeodomain-leucine zipper family I. Role in development and stress responses, IUBMB Life 69 (5) (2017) 280–289, https://doi.org/10.1002/iub.1619.