UNIVERSITY OF
SS. CYRIL AND METHODIUS
IN TRNAVA

Nova Biotechnol Chim (2024) xx(x): e1523
DOI: 10.34135/nbc.1523

Nova Biotechnologica *et* Chimica

Unveiling unknown transcripts and exons: Insights into durian var. D24 (*Durio zibethinus* Murr.) fruit development and ripening

Nurul Arneida Husin[1,2,✉]

[1]*Department of Biotechnology, Faculty of Applied Sciences, AIMST University, Bedong-Semeling Road, Bedong 08100, Semeling, Kedah, Malaysia*
[2]*Biomedical Research Laboratories, Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway 47500, Selangor Darul Ehsan, Malaysia*

✉*Corresponding author:* nurul.arnieda@monash.edu

Article info

Article history:
Received: 10th June 2022
Accepted: 3rd April 2024

Keywords:
CuffDiff
Durio zibethinus
Fruit-pulp tissues
RNA-Seq
Unknown transcripts (UTs)

Abstract

Durian (*Durio zibethinus* Murr.) plant fruits are popular among Southeast Asians. This study aimed to characterise the gene expression during the growth and development of durian fruit at its young, mature, and ripening stages. We used a high-throughput RNA Sequencing approach to identify and characterise unknown transcripts (UTs) from durian fruit pulp transcriptomes. The categorisation of UTs relied on assessing their coding potential and conducting sequence analysis, followed by thorough manual curation. After mapping, 110 million high-quality reads were analysed for each of the nine samples, revealing that each contains 76,700 to 89,117 distinct transcripts and 561,211 to 646,291 exons. In the present analysis, we emphasised identifying and classifying UTs, disregarding transcripts with accurate annotations. Differential expression analysis identified 280 significant unknown transcripts, presumably involved in various biological functions in durian growth. The top BLASTn comparative analysis results for 183 unknown transcripts (UTs) primarily showed significant homology, ranging from 80 % to 100 %, with transcripts from closely related species within the *Malvaceae* family, including *Bombax ceiba*, *Gossypium hirsutum*, *Gossypium raimondii*, *Herrania umbratica*, and *Theobroma cacao*. Functional annotation showed that many upregulated UTs may encode protein-coding genes involved in cellular and metabolic activities, catalytic and electron transfer activities, cellular and anatomical entities, and protein-containing complexes. Out of 97 UTs that do not have a BLASTn hit, 2 of them match GO terms, TCONS 00034019 and TCONS 00058246, corresponding to the InterPro GO Names, namely P: positive regulation of organ growth (GO:0046622) and F: calcium ion binding (GO:0005509). Three open reading frame (ORF) sequences longer than 300 nucleotides were identified as potential protein-coding genes through SMARTBLAST analysis, showing sequence similarities with Retropepsins, pepsin-like aspartate proteases, DUF4492 domain-containing protein, and a hypothetical protein CTI12_AA006750 in UniProtKB/Swiss-Prot. In conclusion, this study sheds light on the dynamic gene expression patterns during durian fruit development. It highlights the significance of characterising unknown transcripts and their potential roles in biological processes, thus enhancing our understanding of durian genetics.

Introduction

Durian (*Durio zibethinus* Murr.), renowned for its distinctive flavour and aroma, stands as a prized tropical fruit tree. Given its significant economic and cultural significance, there is a pressing need to broaden durian's genomic resources. More research is required to fully understand the genetic basis of its traits and potential for crop improvement. An in-depth understanding of durian fruit molecular biology can help to improve its traits. In addition, prolonging the short post-harvest life of the fruit may be considered for further genetic improvement of the durian (Husin *et al.* 2018). Sulphur-containing compounds cause a strong and pungent durian scent, while the fruity smell is caused by esters and alcohol (Siriphanich 2011). Combining positive traits, fewer odours, and high-stress tolerance would allow one to obtain superior durian clones.

High-throughput sequencing technologies are commonly employed for comprehensive transcriptome profiling due to their ability to offer intricate insights into transcript abundances. Several scientific studies have suggested that deep sequencing can detect differential expression of genes, novel genes, transcripts, differentially expressed isoforms and splice variation unique for each condition (Jakhesara *et al*. 2013). Transcriptome data of Malaysia and Thailand durian (Teh *et al.* 2017; Husin *et al.* 2022; Nawae *et al.* 2023) and associated annotation data may fill in some gaps in durian molecular biology, often focused on phytochemical compositions and nutritional properties. The reference assembly of the durian genome (MK) used in this study consisted of chromosome-scale pseudomolecules with a length of 30 pseudomolecules larger than 10 Mb and comprising 95 % of the 712-Mb assembly (the pseudomolecules are hereafter referred to as chromosomes, numbered by scale). These numbers closely correspond to previous estimates of number of haploid chromosomes in durian (1n = 28, 2n = 56) (Teh *et al.* 2017). Recent research on durian genomes originated in Thailand highlights significant genomic variations among cultivars. The study reveals that a large portion of the Kradumthong (KD), Monthong (MT), and Puangmanee (PM) assemblies align closely with the Musang King (MK) assembly, surpassing 50 % identity. However, certain regions, particularly within repeat sections of the MK assembly, show lower identity percentages (Nawae *et al.* 2023). Unknown transcripts (UTs) can include both coding and non-coding RNA sequences. Therefore, while some UTs may be non-coding RNAs (ncRNAs), others may represent novel protein-coding genes or alternative isoforms of known genes (Zhao *et al.* 2022). While the genomes and transcriptomes of durian of the Malaysian MK variety (Teh *et al.* 2017), Malaysian D24 variety (Husin *et al.* 2022), and three Thailand durian cultivars (Nawae *et al.* 2023) have been extensively studied, little attention has been given to the identification of unknown genes or transcripts expressed in the durian transcriptome. Novel transcripts from durian transcriptome data are pieces of RNA sequences found during a transcriptome analysis but whose origin or functions still need to be discovered. These unknown transcripts are identified by comparing them to known durian transcripts and gene databases. Understanding unknown transcripts is critical for discovering new genes and their roles and having a more comprehensive understanding of the transcriptome. Characterisation of novel transcripts and exons involves identifying and describing these sequences regarding their function and expression patterns. By doing so, we can better understand a particular plant tissue's genetic makeup and its biological processes. Noncoding RNAs (nRNAs) are the name given to these RNAs that are not translated into proteins. Although non-coding RNA (ncRNAs) have been known for some time, their inadequate annotation makes much of the RNA-seq data difficult to interpret (Weirick *et al.* 2016). Even though only a few RNAs are translated into proteins, the non-coding RNAs play critical roles in various biological processes, including gene regulation, stress response, and development. Thus, understanding the functions and mechanisms of these non-coding RNAs is an essential area of research in plant biology. In eukaryotic cells, including plant cells, most of the genome is transcribed into RNA, but only a tiny proportion of those RNAs are translated into proteins. This is because many transcribed RNA molecules serve non-coding functions such as regulatory roles, processing of other RNAs, or structural roles rather than encoding proteins. In plants, some of the non-coding RNAs that have been identified include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), and small nuclear RNAs (snRNAs) (Yu *et al.* 2019). Other types of non-coding RNAs that have been found in plants include microRNAs (miRNAs), long non-coding RNAs (lncRNAs), and circular RNAs (circRNAs) (Kung *et al.* 2013; Chao *et al.* 2022). Plant non-coding RNAs (ncRNAs) and microRNAs (miRNAs) significantly influence several plant biological processes, including regulating their gene expression (Yang *et al.* 2020).

This study's characterisation of unknown transcripts and exons likely utilised techniques such as RNA sequencing, gene expression analysis, and functional annotation to identify previously unannotated transcripts and exons, providing insights into their potential functions and biological roles. Enhancing the annotation of durian fruit pulp transcripts contributes to a better understanding of the genetic and molecular mechanisms underlying its development, maturation, and nutritional and sensory properties. It can inform the development of new durian varieties with enhanced traits and help identify potential targets for crop improvement and breeding programs.

Our paper reports identifying and characterising unknown transcripts and exons from the transcriptome of durian (D24 var.) fruit pulp in its three stages of development (young, mature, and ripening). These RNA-unknown transcripts have been detected in experimental transcriptome data of the D24 durian variety (Husin *et al.* 2022) but have not been previously annotated or characterised. The study is expected to discover new transcripts and exons that will aid in the genetic enhancement of durian and improve the annotation of the published draft of the durian genome (Teh *et al.* 2017). The results of this transcriptomics research can be used to investigate further and unknown genes for potential and commercial uses. The characterisation of novel transcripts and exons in durian fruit pulp can contribute significantly to our understanding of this vital crop and serve as a valuable resource for future research.

Experimental

Plant material

Fruit pulp samples were collected from three different stages (young, mature, and ripening) of durian variety (clone D24) fruits at the Agriculture Park of Universiti Putra Malaysia (UPM), Selangor, Malaysia. Three biological replicates of the durian variety clone D24 fruits were harvested at 90 days. The mature fruit (120 days) was left to ripen naturally at room temperature for seven days, resulting in a total age of 127 days, to obtain pulp tissue samples during the ripening stage.

RNA isolation, cDNA library construction and HiSeq Illumina sequencing

The RNA extraction kit, named the GeneAll RibospinTM Seed/Fruit RNA mini kit (Geneall

Biotechnology Co., Ltd), was used to separate the RNA. Using the NEB Next Poly(A) mRNA isolation module, poly-A mRNAs were purified from 1 µg of total RNA. The first and second strands of cDNA synthesis were conducted in the NEB Next RNA Library Prep kit with purified mRNA (Illumina). All protocols were performed according to the manufacturer's instructions. Before sending it for sequencing, the Qubit 2.0 Fluorometer and Agilent Tape Station measured the library's consistency and volume. Nine pre-prepared cDNA libraries have been submitted to the NGS service provider (Novogene, China). The libraries were sequenced using a 2 × 150 bp pair-end protocol with the HiSeq Illumina Platform. The data output predicted was 16.7 million reads or 5 GB per sample. The total data for the nine samples was 150.3 M or 45 GB.

Pre-processing of Illumina RNA-Seq reads using FastQC and Trimmomatic

FastQC is a Java software used to check the quality of the transcriptome. The purpose of running FastQC is to determine the quality encoding of fastq files, to assess the quality of sequencing samples, and to identify overrepresented sequences-adapters and other potential contamination that may occur in the sequencing process. FastQC was run to check the quality of the raw data and the quality of the trimmed data. The comparison was made before and after (twice for each sample) to ensure only the cleaned reads were used for downstream analysis (Andrews 2010).

Trimmomatic is a tool that performs a variety of trimming of the Illumina FastQC of paired-end or single-end data. It removes the added adapters during sequencing (Blankenberg *et al.* 2010). The parameter settings in trimmomatics are as follows: PE = Paired-end (two separate files), ILLUMINACLIP = true, standard adapter, max mismatch = 2, SLIDINGWINDOW = number of bases to average across = 1, average quality required = 20, HEADCROP = 15, nine samples. The distance function was applied to the transpose of the transformed count matrix to get the sample-to-sample distances.

Mapping with TopHat

TopHat is an efficient tool for mapping splice junctions in RNA-Seq data. It uses the ultra-high-throughput short-read aligner Bowtie to align RNA-Seq reads to mammalian-sized genomes. It then analyses the mapping findings to detect splice junctions between exons. The default options for TopHat were utilised: ‘20’ for maximal alignments, ‘2’ for final read mismatches, and ‘yes’ for the BAM output of unmapped reads. The output of TopHat is a set of alignments and junctions in the form of a BAM file. This file can be further processed using other tools, such as Cufflinks, to identify transcripts and quantify gene expression levels.

Identification of unknown transcripts using CufDiff

Cufflinks packages include Cuffmerge, Cuffcompare, and CuffDiff. Cufflink assembles transcripts, calculates their abundance, and tests in RNA-Seq samples for differential expression and function. It recognises, reads, and aligns RNA-Seq into an economical collection of transcripts. Based on how many reads support each one, Cufflinks estimates these transcripts' relative abundances (Blankenberg *et al*. 2010). CuffMerge was run to merge two or more transcript assemblies; the output is the merged transcriptome. Cufflinks also include Cuffdiff, which accepts the reads assembled from two or more biological conditions and analyses their differential expression of genes and transcripts, thus aiding in investigating their transcriptional and post-transcriptional regulation under different conditions. Data analysis was conducted using CuffDiff to obtain the list of differentially expressed genes. Cuffdiff finds significant changes in transcript expression, splicing, and promoter use. The results have reported the list of differentially expressed genes.

Gene and transcript expression analyses were studied statistically. The up-regulated and down-regulated genes and transcripts were ranked and determined using the q value and fold change. FPKM (Fragments per kilobase of transcript per million mapped fragments) value was used to evaluate the abundance of the genes in the differential group. If the absolute value of the expression log-2-fold change is more significant than 1.5 and the false discovery rate (FDR)-corrected p-value is less than 0.05, the differential gene expression is statistically significant. A series of analyses were determined, such as the number of expressed genes/transcripts, the number of up and down-regulated genes/transcripts, and the number of differentially expressed genes/transcripts.

Fig. 1 shows the schematic diagram of the entire study workflow. The input dataset was uploaded to FastQC to check the quality. Then, we will continue running trimmomatic and fast QC to check the quality after trimming. After pre-processing and trimming, TopHat was used to align the reads to the Durian Reference Genome. TopHat reads RNA-Seq data and serves as a rapid splice junction mapper. It first aligns the RNA-Seq readings and then analyses the mapping results to locate splice junctions between exons. The transcripts that corresponded to the reference genome were assembled using Cufflink. Cuffdiff was then used to analyse differential expression. Additionally, mapping findings were viewed locally using the Integrative Genomics Viewer (IGV).

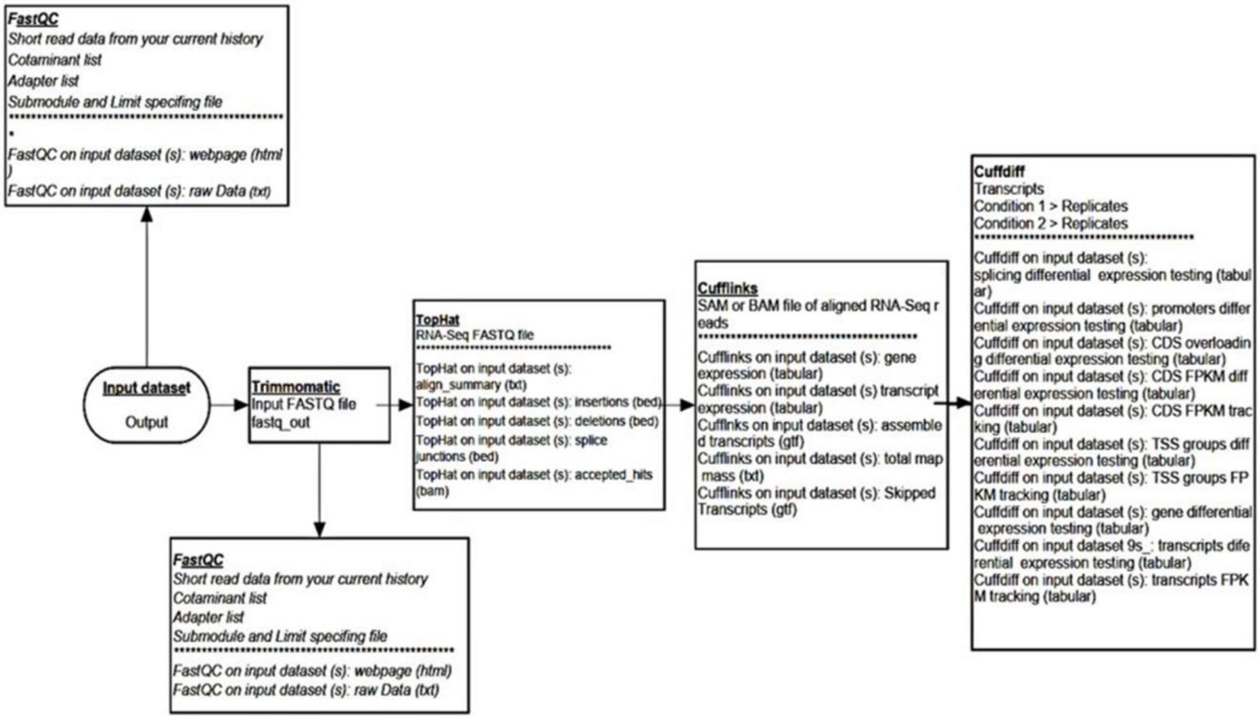
FastQC
Short read data from your current history
Cotaminant list
Adapter list
Submodule and Limit specifing file
FastQC on input dataset (s): webpage (html)
FastQC on input dataset (s): raw Data (txt)
Input dataset
Output
Trimmomatic
Input FASTQ file
fastq_out
FastQC
Short read data from your current history
Cotaminant list
Adapter list
Submodule and Limit specifing file
FastQC on input dataset (s): webpage (html)
FastQC on input dataset (s): raw Data (txt)
TopHat
RNA-Seq FASTQ file
TopHat on input dataset (s): align_summary (txt)
TopHat on input dataset (s): insertions (bed)
TopHat on input dataset (s): deletions (bed)
TopHat on input dataset (s): splice junctions (bed)
TopHat on input dataset (s): accepted_hits (bam)
Cufflinks
SAM or BAM file of aligned RNA-Seq reads
Cufflinks on input dataset (s): gene expression (tabular)
Cufflinks on input dataset (s) transcript expression (tabular)
Cufflnks on input dataset (s): assembled transcripts (gtf)
Cufflinks on input dataset (s): total map mass (txt)
Cufflinks on input dataset (s): Skipped Transcripts (gtf)
Cuffdiff
Transcripts
Condition 1 > Replicates
Condition 2 > Replicates
Cuffdiff on input dataset (s): splicing differential expression testing (tabular)
Cuffdiff on input dataset (s): promoters differential expression testing (tabular)
Cuffdiff on input dataset (s): CDS overloading differential expression testing (tabular)
Cuffdiff on input dataset (s): CDS FPKM differential expression testing (tabular)
Cuffdiff on input dataset (s): CDS FPKM tracking (tabular)
Cuffdiff on input dataset (s): TSS groups differential expression testing (tabular)
Cuffdiff on input dataset (s): TSS groups FPKM tracking (tabular)
Cuffdiff on input dataset (s): gene differential expression testing (tabular)
Cuffdiff on input dataset 9s_: transcripts diferential expression testing (tabular)
Cuffdiff on input dataset (s): transcripts FPKM tracking (tabular)

Fig. 1. Workflow Schematic of the Study. The diagram illustrates the stepwise workflow employed in the study. Beginning with the quality assessment using FastQC, the input dataset undergoes pre-processing and trimming through trimmomatic and FastQC. Next, TopHat, a rapid splice junction mapper, aligns the reads to the Durian Reference Genome, identifying splice junctions between exons. Subsequently, Cufflink assembles transcripts corresponding to the reference genome and Cuffdiff analyses differential expressions.

Comparison to reference annotation

The Illumina HiSeq 4000 RNA-Seq Data of *Durio zibethinus* (GSE136290) was used to improve the transcript annotation of the durian genome. The Cuffcompare program from Cufflinks was used to compare assembled merged transcript assemblies to annotated transcripts, which reveals novel exons and loci. This program uses genomic coordinates to generate a set of transcripts that may be used to identify unknown and well-known transcripts from the chosen reference annotation.

The input for cuff-compare was merged GTF files generated by Cufflinks with a *D. zibethinus* reference annotation file. The output of cuff-compare was transcript accuracy files, which report various statistics related to the accuracy of the transcripts in each sample compared to the reference annotation files. The 'transcript combined files' were written as a GTF file containing each sample's union of transfrags. If a transfrag is present in both samples, it is thus reported once in the combined gtf. This ensures that the merged GTF file contains a comprehensive and non-redundant set of transcripts that captures the full complexity of gene expression in the samples. The 'transcript tracking file' matches transcripts up between samples. The transcript IDs can be used to identify the corresponding novel transcripts in the GTF file generated by Cuffcompare. The Cuffdiff tool was used to analyse differential gene expression using the GTF file containing known and unknown transcripts.

The output of Cuffdiff was examined to identify the significantly differentially expressed transcripts. Cuffcompare and Cuffdiff were run with default parameters.

BLASTn was also used to compare the unknown sequences to the known annotated sequences in the NCBI Genebank database (The National Center for Biotechnology Information) (Syngai *et al.* 2013). All the possible novel transcripts were viewed and downloaded using IGV (Integrated Genome Viewer) for nucleotide search. The FASTA sequences of unknown transcripts were uploaded and used as a query for a BLAST search in NCBI. From the BLASTn search result (E-value <1.0E^{-3}), the transcripts annotated to the known protein-coding gene of other organisms were identified.

After BLAST analysis, the public EMBL-EDI InterPro web-service database was used to analyse the unknown transcript sequences (with no hits in BLASTn) against InterPro's signatures. The unknown sequence information (FASTA) was uploaded to the InterPro server to initiate the InterPro scan. Nucleotides were translated to amino acids automatically in the system. The InterPro scan is a collection of protein families, domains and signals of peptides. Selected databases available in the analysis were identified from PHOBIUS, TMHMM, GENE3D, PROSITE_PATTERNS, SUPERFAMILY, and PANTHER. Fig. 2 shows the workflow analysis of unknown transcript in durian transcriptomics.

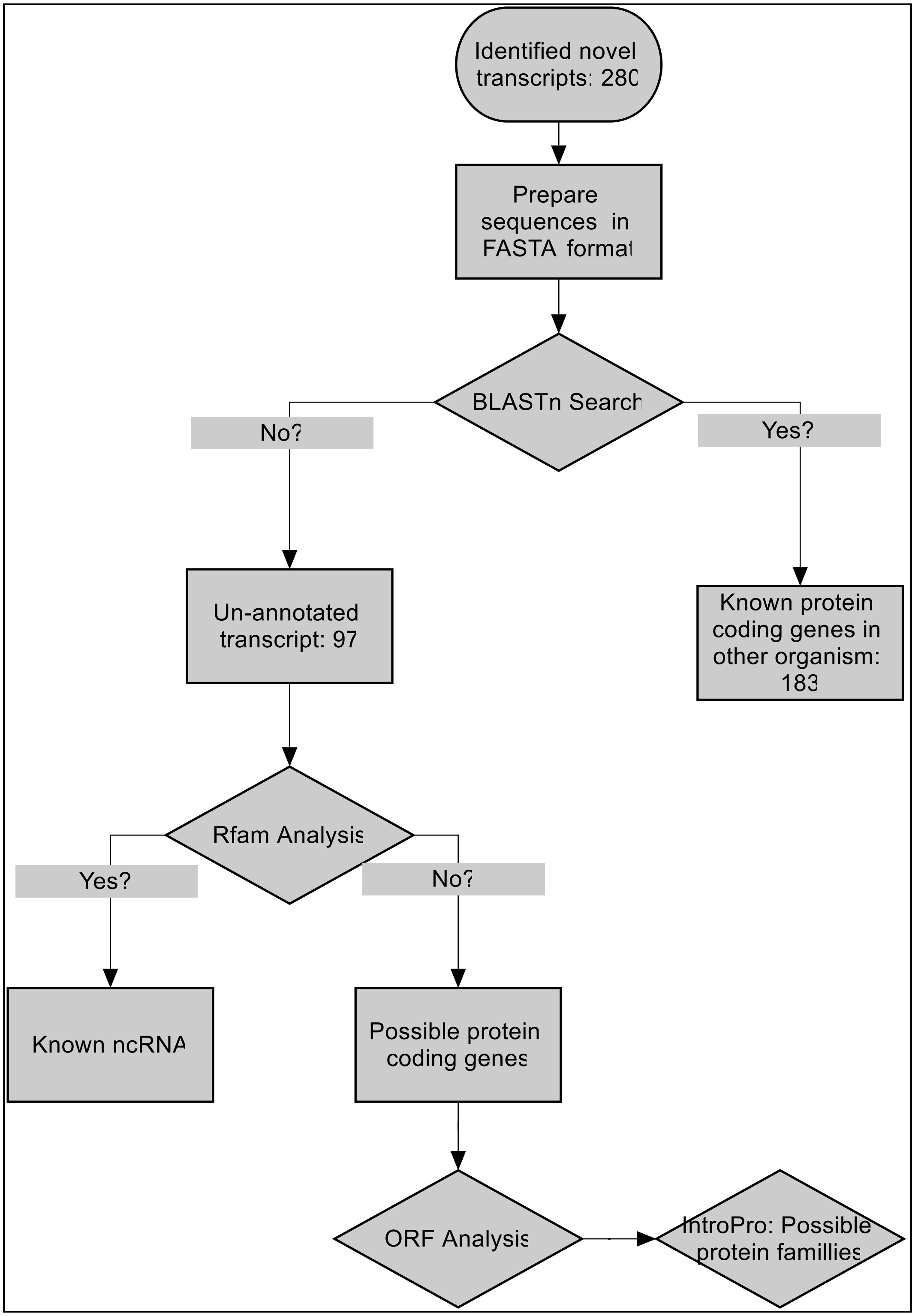
Identified novel transcripts: 280
Prepare sequences in FASTA format
BLASTn Search
No?
Yes?
Un-annotated transcript: 97
Known protein coding genes in other organism: 183
Rfam Analysis
Yes?
No?
Known ncRNA
Possible protein coding genes
ORF Analysis
IntroPro: Possible protein famillies

Fig. 2. Workflow analysis of unknown transcripts in durian transcriptomics. Following BLAST analysis, unknown transcript sequences (with no hits in BLASTn) were analysed using the public EMBL-EDI InterPro web service database. The InterPro scan, encompassing protein families, domains, and signal peptides, was initiated by uploading the unknown sequence information (FASTA) to the InterPro server. Nucleotides were automatically translated to amino acids in the system. The analysis utilised selected databases from PHOBIUS, TMHMM, GENE3D, PROSITE_PATTERNS, SUPERFAMILY, and PANTHER for comprehensive characterisation.

Identification of potential protein-coding regions

A sequential BlastX methodology was employed to identify potential protein-coding genes from unknown transcripts (UTs). Initially, UTs sequences exhibiting hits in BLASTn from the transcriptome data were extracted. Subsequently, these sequences were translated into amino acid sequences using standard genetic code tables, generating all possible reading frames. A protein sequence database for comparison was compiled, comprising a comprehensive collection of known proteins from various organisms, notably the NCBI non-redundant (nr) protein database. The translated amino acid sequences of the UTs served as query sequences in BlastX analyses, with the protein sequence database utilised as the target database. During the BlastX analysis, specific parameters such as the E-value threshold, bit score cutoff, and alignment length were adjusted to accurately identify significant alignments between the translated amino acid sequences of the UTs and the protein sequences in the database. Regions within the UTs displaying substantial similarity to known protein-coding genes in the database were identified as potential protein-coding regions, with lower E-values ($1.0E^{-3}$) and higher bit scores indicating more significant potential.

Results and Discussion

This research conducts transcriptome analysis through deep RNA sequencing to investigate gene transcription in durian fruit across three stages of maturation: young, mature, and ripe. We focus on identifying and categorising unknown transcripts while disregarding transcripts with precise annotations. The study illustrates the extensive potential of RNA Seq in durian for annotating novel transcripts from the *D. zibethinus* genome. We discovered previously unidentified transcripts and exons exhibiting differential expression patterns, potentially influencing durian growth and development.

Heat maps are effective visualisation tools in expression analysis studies spanning diverse biological domains. In this study, heat maps were generated for each of the three groups using transformed data to examine similarities and differences among samples. Additionally, sample-level quality control (QC) was performed to assess how well replicates clustered together. Also, the test reveals if any sample outliers should be removed before analysing DE. Fig. 3A shows the gene expression correlation between young and mature levels. The hierarchical tree indicates that the clustered samples are similar. The dark blue colour blocks represent the substructure in the mature stage sample, replicating data, and high similarities. When the young-stage duplicates are compared, they have fewer similarities. However, the duplicates in the early stages were grouped and well-suited for DE analysis. The relationship between gene expression at the young and ripening stages is shown in Fig. 3B. The sample replicates from the ripening stage have a high degree of similarity and are grouped. Replicates from the early stages have fewer similarities but are still grouped. Fig. 3C shows the gene expression correlation between the mature and ripening stages. This correlation showed similar sample replicates from their group's mature and ripening stages. All blocks in light blue indicate the dissimilarities of the combination of groups. All samples were suitable to proceed with the differential analysis.

Nine samples collected from three biological replicates of *D. zibethinus* from different growth stages were subjected to RNA-Seq using TopHat, Cufflinks, Cuffmerge and Cuffdiff tools. The paired-end reads' sequence was aligned to the *D. zibethinus* reference genome using TopHat (Galaxy Version 2.1.1). This tool accepts files in Sanger FASTQ format. TopHat has produced four output files:

junctions (BED track of junction), insertions, deletions and accepted_hits (BAM format of alignment files). The tool has been chosen as a mapping tool because it can create a database of splice junctions based on the model gene's annotation (Trapnell *et al.* 2012).

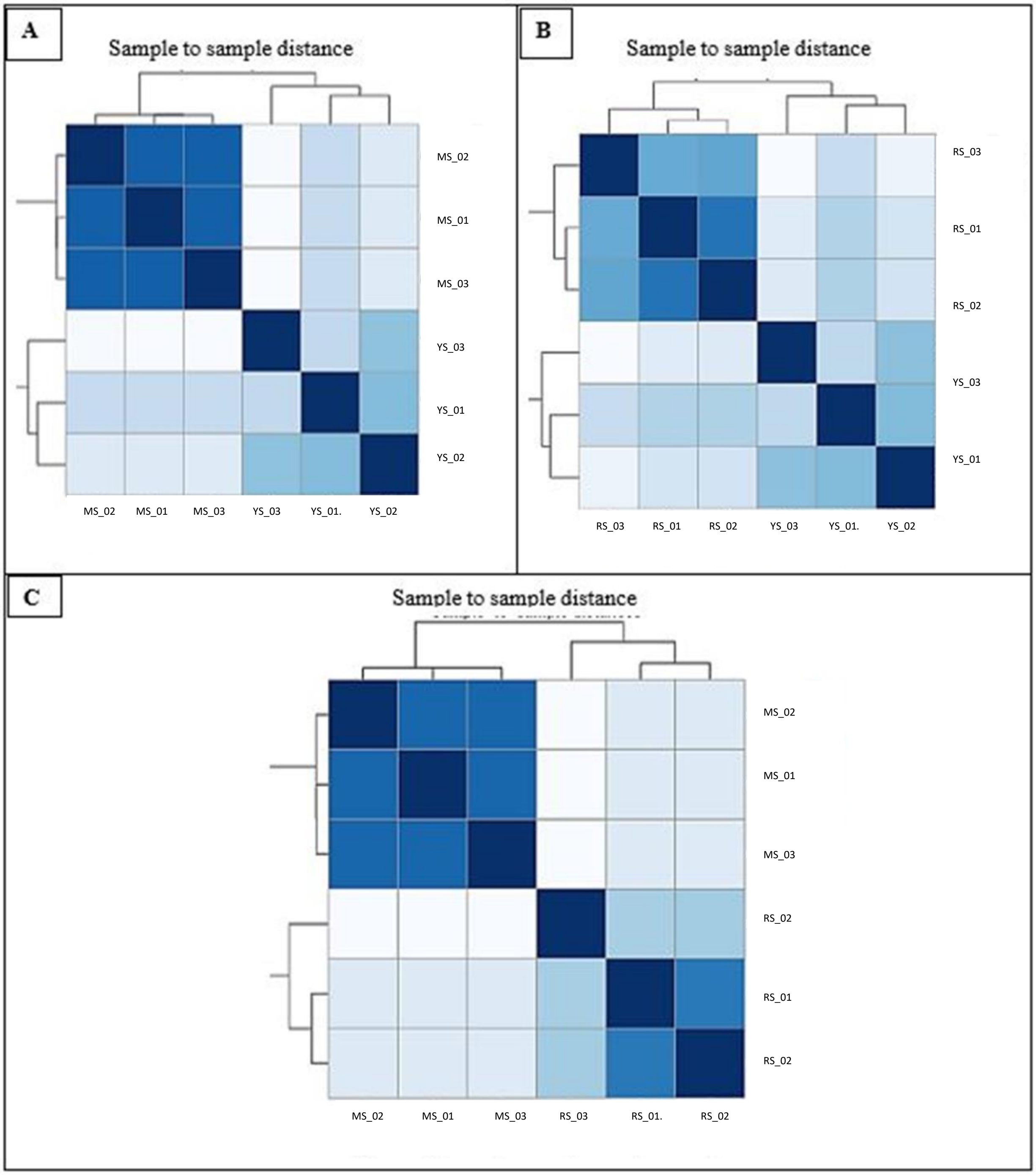

Fig. 3. Heatmap of the sample-to-sample distances using transformed data. A. Young Stage vs Mature Stage (YS/MS), B. Young Stage vs Ripening Stage (YS/RS), and C. Mature Stage vs Ripening Stage (MS/RS).

Table 1 shows mapping results against *Durio zibethinus* Reference Genome (Reference Guided Assembly) using TopHat Galaxy ver. 2.1. Cufflinks were used to assemble the transcripts, estimate their abundance using the FPKM value and perform the differential expression testing. Cufflinks were utilised to reconstruct the transcripts together with TopHat. This exon-guided assembly helped capture the differentially expressed genes in samples with well-annotated genomes.

Table 1 Mapping results against *Durio zibethinus* Reference Genome (Reference Guided Assembly) using TopHat Galaxy ver. 2.1.

Samples	**Left reads**		**Right reads**		**Overall Read Mapping Rate [%]**	**Concordant Pair Alignment Rate [%]**
	Input	Mapped	Input	Mapped		
YS_01 (90 dpa)	10,704,730	9,528143	10,704,730	9,495,383	88.	84.6
YS_02 (90 dpa)	12,838,126	11,590,153	12,838,126	11,543,504	90	85.7
YS_03 (90 dpa)	14,362,534	12,879,555	14,362,534	12,846,945	89.6	85.4
MS_01 (120 dpa)	11,887,099	10,662,256	11,887,099	10,616,497	89.5	84.8
MS_02 (120 dpa)	10,816, 378	9,766,055	10,816,378	9,725,422	90.1	85.8
MS_03 (120 dpa)	12,071,611	10,956,503	12,071,611	10,914,407	90.6	86.4
RS_01 (127 dpa)	11,612,398	10,015,743	11,612,398	9,987,238	86.1	82.9
RS_02 (127 dpa)	11,099,558	9,522,929	11,099,558	9,489,748	85.6	82.4
RS_03 (127 dpa)	10,574,544	8,960,631	10,574,544	8,920,718	84	80.9

dpa: days post anthesis

One hundred ten million high-quality reads were analysed for all nine samples, individually identifying 76,700 – 89,117 transcripts and 561,211 – 646,292 exons after the mapping process. Cuffcompare was used to compare the cufflink assemblies to reference annotation files. All reconstructed transcripts were compared by aligning to the Durian Reference Genome's annotation to identify unknown transcripts from known ones. As a result, 3,438 novel exons and 1,197 novel loci were determined to be expressed in the merged samples. This suggests that there may be substantial variability in gene expression and splicing patterns across the samples. These novel exons and loci may be involved in critical biological processes, and their discovery could provide valuable insights into the regulation of gene expression and the diversity of the transcriptome.

The differential expressed genes (DEGs) between the three stages of durian growth were found using the Cuffdiff tool. This will produce a list of DEGs along with the transcript IDs that go with them. The unknown transcripts differentially expressed in durian samples were found using the Cuffcompare tracking file and the Cuffdiff list of DEGs. Novel transcripts can be located by comparing the transcript IDs in the tracking file with those in the Cuffdiff output file. After identifying the unknown transcripts of interest, a BLAST analysis was conducted to determine their function and evolutionary links by comparing them with known sequences from public databases.

Differentially expressed genes and transcripts of known transcripts have been published in (Husin *et al.* 2022) and will not be discussed further here. Our study focuses on the 280 unknown transcripts (Table S1) for further analysis. 183 (Table S2) unknown transcripts were annotated with known protein-coding genes of other plant species. It is considered novel and related to the first found in *D. zibethinus*. Of 280 identified novel transcripts, 183 hit the Blastn for homology search (16 and

63 are up-regulated genes, 83 and 21 are down-regulated genes in YS/MS and YS/RS. An unknown transcript of 97 (21 and 16 up-regulated genes and 31 and 29 down-regulated genes in YS/MS and YS/RS) (Table S3) did not hit any Blastn and was subjected to further analysis to search for protein domains. Seventy (see Table S4) unknown transcripts with no Blastn hit matched the identifier from Interpro annotation. Twenty-seven novel transcripts were categorised under "No IPS match". Out of 27 (Table S5), three unknown transcripts with ORF sequences with more than 300 nt were identified, and 24 with less than 300 nt were considered small ORFs. The sequences of the novel transcripts are well aligned with the published durian reference genome, but they remain unannotated. These necessary transcripts might be involved in different molecular pathways or networks in durian. The transcripts with no hits in BLAST and no IPS match were shortlisted and further analysed using Rfam and ORF. The Rfam database was used to characterise sequences with no similarity to any proteins and identify homologues of known ncRNA families. ORF Finder is a program available on the NCBI website. The program searches for the open reading frame (ORF) for all possible protein-coding regions in the sequences (Cheng *et al.* 2008; Yazhini 2018).

Comparative analysis of unknown transcript to other plant species

BLASTn was used to compare the unknown sequences to the known annotated sequences in the NCBI Genebank database (The National Center for Biotechnology Information). All the possible novel transcripts were viewed and downloaded using IGV (Integrated Genome Viewer) for nucleotide search. The FASTA sequences of unknown transcripts were uploaded and used as queries for a BLAST search in NCBI. From the BLASTn search result, the transcripts that were annotated to the known protein-coding gene of other organisms were identified.

Two hundred eighty unknown transcripts with multiple and single exon structures were identified and extracted from DEG analysis. The size distribution of new novel transcripts ranged from 150 bp to 8 kb, as illustrated in Fig. 4. These unknown transcripts were searched through blast2GO (blastn) to see whether they could match any taxa other than *D. zibethinus* and to investigate their possible novelties.

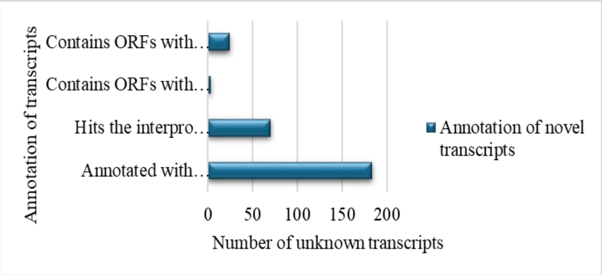

Fig. 4. Characterization of Unknown Transcripts in *Durio zibethinus*. DEG analysis isolated two hundred eighty unidentified transcripts featuring multiple and single exon structures. These transcripts are potentially unknown significant transcripts absent in the Durian Reference Genome. The size distribution of these novel transcripts spans from 150 bp to 8 kb, as depicted in the figure. Subsequent analysis using blast2GO (blastn) explored their potential matches with taxa other than *D. zibethinus* and investigated their possible novelties.

Of 280 identified unknown transcripts, 183 (16 and 63 are up-regulated genes, and 83 and 21 are down-regulated genes in YS/MS and YS/RS) hit the Blastn for homology search of other plant species. Thus, 183 unknown transcripts were considered novel transcripts first found and discovered in *Durio zibethinus*, while 97 were subjected to further analysis to search for protein domains. Since the priority of the study is to recognise only significant transcripts, only differentially expressed unknown transcripts were filtered for further analysis. In the analysis, 183 unknown transcripts were annotated to the known protein-coding gene of other plant species closely related to the durian genome. These annotated transcripts were known as novel-related and were first discovered in durian.

Top blast results (Fig. 5) were showed a homology (80 – 100 %) to *Bombax ceiba*, *Gossypium hirsutum*, *Gossypium raimondii*, *Herrania umbratica*, *Juglans regia*, *Theobroma cacao*, and many other plant species were indicated further in the list.

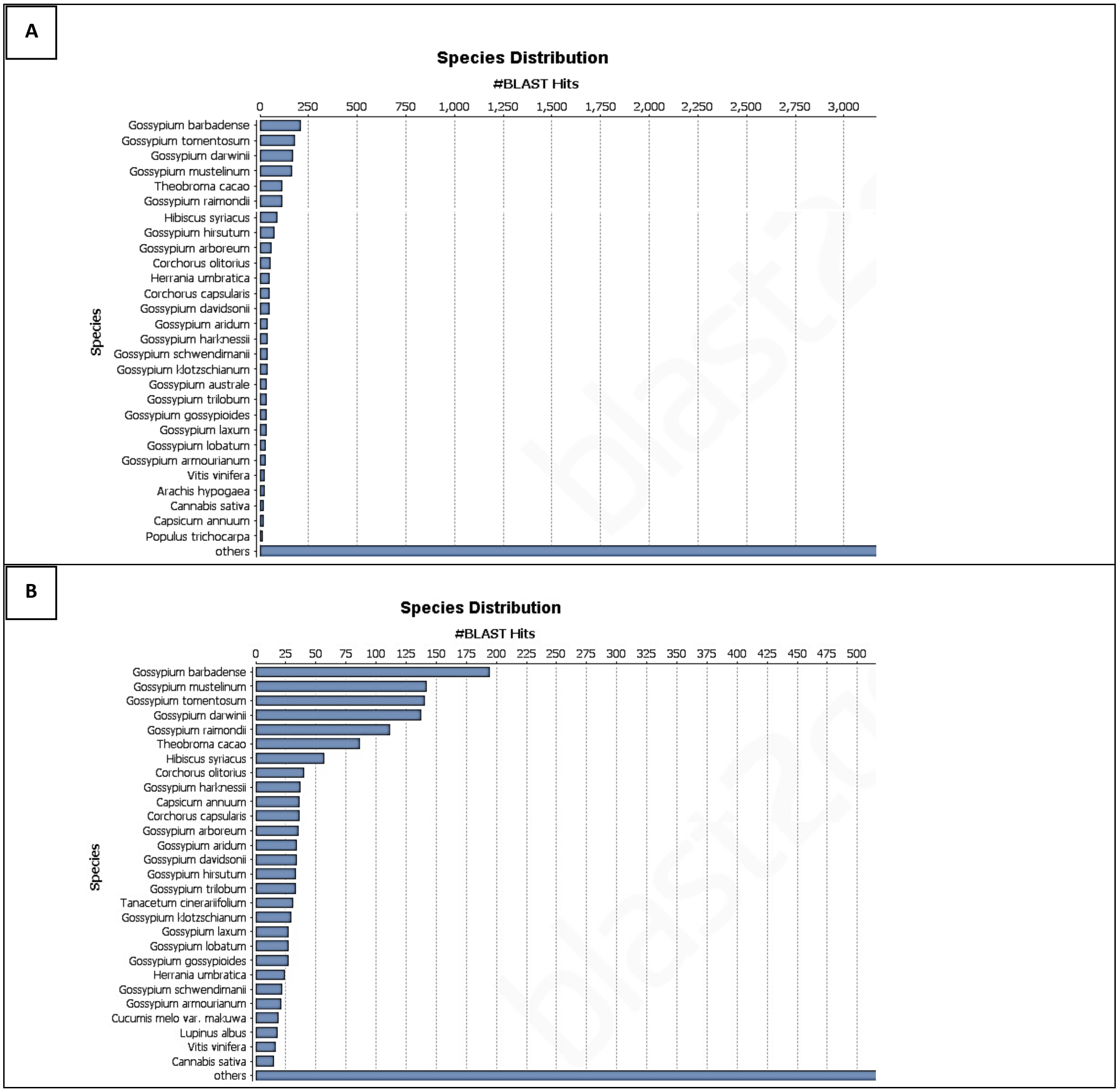

Fig. 5. Distribution of BLAST results for unknown transcripts by species is shown as the percentage of the total homologous sequences (with an E-value <1.0E^{-3}). All plant proteins in the NCBI nr database were used for homology search, and the best hit of each sequence was used for analysis. (A) Young Stage/Mature Stage (YS/MS), (B) Young Stage/Ripening Stage (YS/RS).

The comparison between Fig. 5A and 5B reveals intriguing insights into the expression patterns of the unknown transcripts across different plant developmental stages. The observation that Fig. 5B, representing the ripening stage, demonstrated more BLAST hits of novel transcripts than the development stage suggests potential roles for these genes in late-stage fruit development processes. This highlights the importance of studying gene expression dynamics throughout the developmental process to fully understand the regulatory networks underlying fruit ripening. The top BLASTn results, as presented in Fig. 5, were matched with previous published findings that showed *Theobroma cacao* (cacao used in chocolate) and members of the *Gossypium* genus (cotton) are the available relative genome, including only the more distantly related cash crops within the more prominent Malvaceae family (Teh *et al.* 2017). Malvaceae, or the mallows, is a family of flowering plants estimated to contain 244 genera with 4,225 known species, with known members of economic importance including okra, cotton, cacao, and durian. The Malvaceae s.l. (hereafter merely "Malvaceae") comprise nine subfamilies: Bombacoideae, Brownlowioideae, Byttnerioideae, Dombeyoideae, Grewioideae, Helicteroideae, Malvoideae, Sterculioideae, Tilioideae. The presence of homologous sequences from species within the Malvaceae family, including Bombax ceiba and Gossypium species, highlights the evolutionary relationships among these plants. The Malvaceae family is known for its diverse array of economically important crops. Identifying conserved sequences across different genera within this family provides insights into the genetic basis of traits such as fruit development and ripening that may have been conserved throughout evolutionary history.

The comparative analysis of unknown transcripts to sequences from diverse plant species contributes to our understanding of gene conservation and divergence across evolutionary lineages. By elucidating the evolutionary history of these genes, we can gain insights into the genetic basis of traits shared among distantly related plant species and identify candidate genes for further investigation in crop improvement efforts.

Functional annotation of unknown transcripts in the developmental and ripening stage

Identifying putative functional genes involves a multifaceted approach aimed at elucidating the roles of unknown transcripts, particularly in fruit development and ripening. Initially, novel transcript sequences are aligned with known sequences from related species. This alignment identifies homologous sequences with similar functions, providing a basis for inferring putative functions for the unknown transcripts based on the annotated functions of their homologs.

Subsequently, Blastx analysis is conducted, running 183 unknown transcript sequences against protein databases to explore potential functions comprehensively. This analysis offers insights into the degree of similarity between the novel transcripts and known proteins, aiding in the inference of their roles in fruit development and ripening processes. The inferred putative functions serve as a starting point for further experimental validation, where techniques such as gene expression analysis and functional assays are employed to confirm the predicted functions and elucidate the roles of the transcripts in specific biological pathways.

In the developmental stage of the fruit, Blastx analysis identified 14 up-regulated unknown transcripts encoding protein-coding regions. Among them, three transcripts, ranging from 582 to 1,064 nucleotides in length, were identified as members of the ABC transporter C family. Refer to Table S7, which underscores the significance of these transcripts in fruit development. The ABC transporter gene plays a pivotal role, facilitating crucial cellular processes in the early development of fruits. Through Gene Ontology (GO) analysis, several fundamental functions associated with these ABC transporter transcripts have been elucidated. They are implicated in transmembrane transport, ATP binding, ATP hydrolysis activity, and ABC-type transporter activity. Their primary localisation is also identified as the cellular membrane, aligning with their role as membrane transport proteins. Furthermore, the enzyme code EC:3.6.1.15, representing nucleoside-triphosphate phosphatase activity, sheds light on the catalytic functions of these ABC transporter transcripts. These enzymes facilitate the translocation of inorganic cations and are associated with various metabolic pathways, including purine metabolism, thiamine metabolism, metabolic pathways, and biosynthesis of secondary metabolites.

Identifying and characterising ABC transporter transcripts during the early stages of fruit development provides valuable insights into their essential roles in cellular transport processes and metabolic pathways. ABC transporter C family member 4 (ABCC4) plays a pivotal role in plant fruit development by facilitating the transport of various molecules across cellular membranes. These transporters belong to the ATP-binding cassette (ABC) superfamily. They are renowned for their involvement in the active transport of a broad spectrum of substrates, encompassing ions, lipids, sugars, and secondary metabolites (Hwang *et al.* 2016; Sun *et al.* 2021). ABCC4, in particular, has been implicated in transporting phytohormones, including auxins and abscisic acid, crucial for fundamental fruit development processes such as cell division, expansion, and ripening (Kang *et al.* 2010). Moreover, ABCC4 may contribute to transport defence compounds, metabolites, and nutrients, thereby enhancing fruit quality and bolstering resistance to both biotic and abiotic stresses (Do *et al.* 2021).

During the ripening stage of fruit, Blastx analysis revealed 50 up-regulated unknown transcripts encoding protein-coding regions. Among them, three transcripts, ranging from 278 to 455 nucleotides in length, were identified as NADH-ubiquinone oxidoreductase chain 1 and NADH-plastoquinone oxidoreductase subunit 6, as indicated in Table S8. The Gene Ontology (GO) analysis for NADH-ubiquinone oxidoreductase chain 1 revealed several functions, including proton motive force-driven ATP synthesis and proton transmembrane transport. It also exhibited ATP binding and proton-transporting ATP synthase activity with a rotational mechanism. Localisation was observed in the mitochondrion and proton-transporting ATP synthase complex, catalytic core F(1), with the enzyme commission (EC) number 7.1.2.2. It is also involved in H(+)-transporting two-sector ATPase and is associated with oxidative phosphorylation, photosynthesis, and metabolic pathways.

Regarding the GO ontology of NADH-plastoquinone oxidoreductase subunit 6, identified functions included the electron transport chain, ATP binding, and NADH dehydrogenase (ubiquinone) activity. Its localisation was observed in the membrane with enzyme commission (EC) numbers EC:7.1.1.2 and EC:1.6.5.2, representing NADH: ubiquinone reductase (H(+)-translocating), NADH dehydrogenase (quinone), and NAD(P)H dehydrogenase (quinone). The identified pathways included ubiquinone and other terpenoid-quinone biosynthesis, metabolic pathways, and biosynthesis of secondary metabolites. NADH-ubiquinone oxidoreductase chain one and NADH-plastoquinone oxidoreductase subunit 6 are essential genes in plant fruit ripening, crucial for energy production, proton transmembrane transport, and metabolic pathways. They participate in proton motive force-driven ATP synthesis and ATP binding, ensuring the availability of energy required for fruit ripening processes. Additionally, their involvement in metabolic pathways such as oxidative phosphorylation, photosynthesis, and biosynthesis of secondary metabolites influences fruit flavour, aroma, colour, and nutritional quality. Their localisation in cellular membranes highlights their role in membrane-associated processes, including ion transport and redox signalling, which are vital for cellular homeostasis and physiological responses during fruit ripening (Li *et al*. 2019).

All selected unknown transcripts with q value > 0.05 were uploaded to Blast2GO to annotate its putative functions assigned into GO term within three main categories (BP, MF and CC). Tables S9 and S10 represent GO analysis of up-regulated unknown transcript in the developmental and ripening stages. GO analysis in the developmental stage revealed that for biological process (BP), major sub-categories were ten transcripts in the cellular process (GO:0009987), seven transcripts in the metabolic process (GO:0008152), four transcripts in localisation (GO:0051179), three transcripts in regulation of biological process (GO:0050789), three transcripts in natural regulation (GO:0065007), two transcripts in response to stimulus (GO:0050896), one transcripts in negative regulation of biological process (GO:0048519), and one transcripts in positive regulation of natural process (GO:0048518). As for molecular function (MF), 4 transcripts in transporter activity (GO:0005215), five transcripts in catalytic activity (GO:0003824), 2 transcripts in structural molecule activity (GO:0005198), 7 transcripts in binding (GO:0005488), 1 transcript in molecular function regulator activity (GO:0098772), and 4 transcripts in ATP-dependent activity (GO:0140657). Cellular component (CC) enriched with 11 transcripts in cellular anatomical entity (GO:0110165) and 3 transcripts in protein-containing complex (GO:0032991).

GO analysis in ripening stage revealed that for biological process (BP), major sub-categories were 21 transcripts in the cellular process (GO:0009987), 7 transcripts in localization (GO:0051179), 20 transcripts in metabolic process (GO:0008152), 2 transcripts in biological regulation (GO:0065007) and 2 transcripts in regulation of biological process (GO:0050789). As for molecular function (MF), 8 transcripts in electron transfer activity (GO:0009055), 17 transcripts in catalytic activity (GO:0003824), 14 transcripts in binding (GO:0005488), 4 transcripts in structural molecule activity (GO:0005198) and 10 transcripts in transporter activity (GO:0005215). 43 transcripts in cellular anatomical entity (GO:0110165) and 12 transcripts in protein-containing complex (GO:0032991) were enriched in cellular components.

Different functional patterns emerge when comparing the GO analysis results between the developmental and ripening stages of durian (*Durio* spp.). Regarding biological processes, the developmental stage is characterized by a balance between various fundamental cellular processes, such as cellular and metabolic activities, and regulatory mechanisms. In contrast, the ripening stage appears to be more focused on metabolic activities, suggesting a shift towards processes related to the maturation and ripening of the durian fruit. Regarding molecular functions, the ripening stage emphasizes catalytic and electron transfer activities, which could be associated with biochemical transformations during ripening. Additionally, the increase in structural molecule activity during this stage suggests potential changes in the structural components of the durian fruit. Regarding cellular components, both developmental and ripening stages exhibit enrichment in cellular anatomical entities, indicating active cellular processes. The significant increase in protein-containing complexes during the ripening stage may reflect protein composition or assembly changes as the durian fruit matures.

Analysis of protein domains using InterPro Scan

A set of 97 transcripts (21 and 16 up-regulated genes and 31 and 29 down-regulated genes in YS/MS and YS/RS) did not produce any hit with Blastn. These 97 novel transcripts did not show any similarity with any sequences available in the gene bank's nucleotide database. The public EMBL-EDI InterPro web service database analysed the 97 unknown transcript sequences (with no hits in BLASTn) against InterPro's signatures. The unknown sequence information (FASTA) was uploaded to the InterPro server to initiate the InterPro scan. Nucleotides were translated to amino acids automatically in the system. As a result, 70 unknown transcripts (Table S4) hit the Interpro annotation, and the other 27 unknown transcripts (Table S5) were categorised under "No IPS match".

Fig. 6 shows the expression analysis of an unknown transcript that hits GO terms using IGV. Under the category of unknown transcripts with no Blastn hit, two unknown transcripts hit GO terms; TCONS_00034019 (Fig. 6A) and TCONS_00058246 (Fig. 6B). TCONS_00034019 [locus position: NW_019167937.1:10860347-10862497] hits the InterPro GO Names for Biological Process (P) P: GO:0046622; P: positive regulation of organ growth. The sequences were extracted from IGV and uploaded to BLASTx to see the match. The top hits for the sequences were identified as the auxin-regulated gene involved in organ size [Theobroma cacao] with a similarity of 92.1 % and E value of 3e-17. TCONS_00058246 [locus position: NW_019168015.1:11786205-11786925] hits the InterPro GO Names for Molecular Function (F) F: GO:0005509; F: calcium ion binding. The sequences were extracted from IGV and uploaded to BLASTx to see the match. The sequences' top hits were identified as an uncharacterized protein TCM_029153 [*Theobroma cacao*], with a similarity of 57.97 % and an E-value of 7e-17. InterPro scan is a collection of protein families, domains and signals of peptides. The selected database in the analysis is identified from PHOBIUS, TMHMM, GENE3D, PROSITE_ PATTERNS, SUPERFAMILY, and PANTHER.

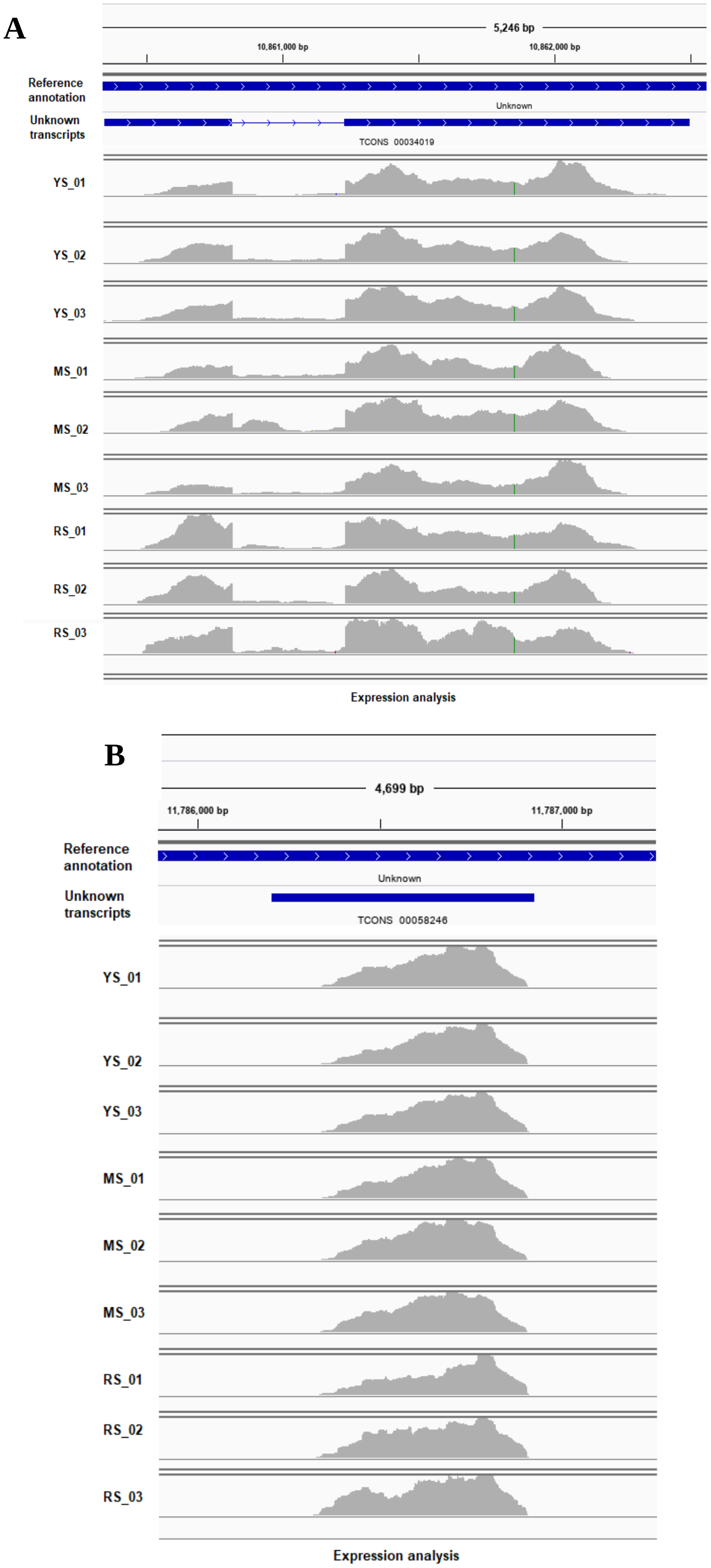

Fig. 6. Expression analysis of unknown transcript that hits GO terms using IGV. **A** – Transcript id of TCONS_00034019. **B** – Transcript id of TCONS_000.

Analysis of Rfam and ORF of unknown sequences with no IPS match

The Rfam database was used to characterise sequences with no similarity to any proteins and identify homologues of known ncRNA families. The RNA families were classified into functional classes: ncRNA genes, structured cis-regulatory elements and self-splicing RNAs. The Rfam database is represented by multiple sequence alignments, consensus secondary structures and covariance models (Cms). The transcripts with no hits in BLAST and no IPS match (Table S6) were shortlisted and further analysed using Rfam. From the results of the Rfam search, the transcripts with hits are known as non-coding genes. Non-coding RNAs analysis can be carried out by constructing the co-expression network to characterise the functions of non-coding RNAs (Kalvari *et al.* 2018). The study of unknown sequences displayed no hits results in the Rfam database, and no further information is available about them.

By searching open-reading frames (ORFs), three transcripts with ORFs of more than 300 nt were found, and twenty-four with less than 300 nt were seen. Comparative sequence analysis using SMARTBLAST refers to a bioinformatics tool called SMART (Simple Modular Architecture Research Tool) that utilizes a specialized version of BLAST (Basic Local Alignment Search Tool) for comparing query sequences against a large database of protein sequences. SMARTBLAST is particularly useful for identifying domains and motifs within protein sequences, providing insights into their structure, function, and evolutionary relationships. The tool was used to compare the sequences of unknown transcripts identified in durian fruit with sequences in the UniProtKB/Swiss-Prot database, helping to infer potential functions and similarities with known proteins.

Under the category of sequences with no BLASTx hit and no IPS match, there were three ORFs with more than 300 nt lengths. The locus XLOC_004344 (consists of TCONS_00009381) with 1,650 nt; 549 aa was classified as an unknown protein-coding gene. This transcript was differentially expressed between the young and mature stages in the up-regulated level and was not annotated in the reference durian genome. Comparative sequence analysis using SMARTBLAST based on UniProtKB/Swiss-Prot detected sequence similarity with DOMAIN: Retropepsins; pepsin-like aspartate proteases uncharacterised protein LOC100803389 [*Glycine max*] with e-value of 2e-06 and identity 23.12 %.

The locus XLOC_035884 and XLOC_008159 (consists of TCONS_00076984 and TCONS_00017315) with 324 nt; 107 aa and 318 nt; 105 aa, were classified as unknown protein-coding genes. These transcripts were differentially expressed between the young and ripening stages in the downregulated level and were not annotated in the reference durian genome. Comparative sequence analysis using SMARTBLAST based on UniProtKB/Swiss-Prot detected sequence similarity with DUF4492 domain-containing protein [*delta proteobacterium NaphS2*] with an e-value of 0.007 and identity of 46.67 % and hypothetical protein CTI12_AA006750 [*Artemisia annua*] with e-value of 9e-11 and identity of 50.0 %.

Validation of UTs

We have employed a simple strategy of cross-verifying unidentified transcripts in BLAST and IGV. This statement implies that two lines of evidence supported the identification of novel transcripts and exons: (1) visual checking of read alignments using the Integrated Genome Viewer (IGV) software and (2) differential expression testing results. The alignment was viewed using the Integrated Genome Viewer (IGV), and results showed the existence of these novel exons. IGV is a well-known genome browser that allows researchers to explore and investigate genomic data in various forms, such as read alignments and variant calls. In this instance, read alignments in IGV, which most likely revealed evidence of previously undetected splice junctions or other traits that suggested the existence of novel transcripts, allowed us to detect the presence of novel exons.

Results from tests for differential expression may have also shown the existence of these new exons. A statistical procedure called differential expression testing examines gene or transcript expression levels in various samples or conditions. This could further support their existence if the novel exons are expressed at significantly different levels in different samples or conditions. To better comprehend the complexity and diversity of gene expression in biological systems, combining IGV and differential expression testing can give complementary lines of evidence for the occurrence of novel exons and other transcript characteristics.

This study has shown that these gaps must be filled to help future RNA-Seq studies of Malaysian durian fruit. While RNA-Seq can detect novel gene structures without prior annotation information, rigorous detection and validation approaches are required to exclude false transcripts, particularly in the case of low coverage and short readings provided by most current NGS technologies. However, software such as Cufflinks (19) is available for transcript reconstruction from RNA-Seq data, which can be used for transcript assembly, given that the sequencing is done with adequate coverage and more extended readings.

One limitation of our study is that we did not validate the presence of the identified unknown transcripts using qPCR. While RNA-seq is a powerful tool for transcriptome analysis, it can sometimes generate false positives. Validation with an independent method like qPCR is essential to confirm the results. With qPCR validation, some identified transcripts may be biologically relevant or present at lower levels than suggested by the RNA-seq data. Therefore, the lack of qPCR validation in our study should be considered a limitation, and the results should be treated with caution until more validation is performed.

Conclusions

The RNA-seq investigation that we performed on durian fruit pulp transcriptome has provided us with insightful new information regarding the genome of this fruit. We can identify unidentified transcripts and exons by employing Illumina sequencing to produce 150 bp reads, which has improved the annotation of the durian draft genome. Our study gives comprehensive information on these hitherto unknown characteristics, which can be used to comprehend the biology of durian and assist the future investigation into this unique fruit. Overall, our work represents a significant contribution to the field of durian genomics and has the potential to advance our understanding of durian and other related species.

Acknowledgements

The author wants to acknowledge that the research funding for this project was provided by AIMST University, Malaysia, under its internal research grant scheme.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

Andrews S (2010) FastQC: A Quality Control tool for High Throughput Sequence Data. Babraham Bioinformatics.

Blankenberg D, Kuster GV, Ananda NCG, Lazarus R, Mangan M, Nekrutenko A, Taylor J (2010) Galaxy, a web-based genome analysis tool for experimentalists Daniel. Curr. Protoc. Mol. Biol. 19:

1-33.

Chao H, Hu Y, Zhao L, Xin S, Ni Q, Zhang P, Chen M (2022) Biogenesis, Function, Interactions, and Resources of Non-Coding RNAs in Plants. INT. J. Mol. Sci. 23: 3695.

Cheng H, Chan WS, Li Z, Wang D, Liu SYZ (2008) Small Open Reading Frames: Current Prediction techniques and Future Prospect. Curr. Protein Pept. Sci. 23: 1-7.

Do THTH, Martinoia E, Lee Y, Hwang J-U (2021) Update on ATP-binding cassette (ABC) transporters: how they meet the needs of plants. American Society of Plant Biologists, Plant Physiol. 187: 1876-1892.

Husin NA, Rahman S, Karunakaran R, Bhore SJ (2018) A review on the nutritional, medicinal, molecular and genome attributes of durian (*Durio zibethinus* L.), the King of fruits in Malaysia. Bioinformation 14: 265-270.

Husin NA, Rahman S, Karunakaran R, Bhore SJ (2022) Transcriptome analysis during fruit developmental stages in durian (*Durio zibethinus* Murr.) var. D24. Genetics Mol. Biol. 45: 4.

Hwang Ja-U, Song Won-Y, Hong D, Ko D, Yamaoka Y, Jang S, Yim S, Lee E, Khare D, Kim K, Palmgren M, Yoon HS, Enrico M, Lee Y (2016) Plant ABC Transporters Enable Many Unique Aspects of a Terrestrial Plant’s Lifestyle. Mol. Plant 9: 338-355.

Jakhesara SJ, Koringa PG, Joshi CG (2013) Identification of novel exons and transcripts by comprehensive RNA-Seq of horn cancer transcriptome in Bos indicus. J. Biotechnol. 165: 37-44.

Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, Bateman A, Petrov AI (2018) Non-Coding RNA Analysis Using the Rfam Database. Curr. Protoc. Bioinformat. 62: 1-27.

Kang J, Hwang Jae-U, Lee M, Kim Yu-Y, Assmann SM, Martinoia E, Lee Y (2010) PDR-type ABC transporter mediates cellular uptake of the phytohormone abscisic acid. PNAS 107: 2355-2360.

Kung JTY, Colohnori D, Lee JT (2013) Long Noncoding RNAs: Past, Present and Future. Genetics Society of America.

Li X, Liu L, Ming M, Hu H, Zhang M, Fan J, Song B, Zhang S, Wu J (2019) Comparative Transcriptomic Analysis Provides Insight into the Domestication and Improvement of Pear (*P. pyrifolia*) Fruit. Plant Physiol. 180: 435-452.

Nawae W, Naktang C, Charoensri S, U-thoomporn S, Narong N, Chusri O, Tangphatsornruang S, Pootakham W (2023) Resequencing of durian genomes reveals large genetic variations among different cultivars. Front. Plant Sci. 14: 1137077.

Siriphanich J (2011) Durian (*Durio zibethinus* Murr.), in: *Durio Zibethinus* Murr. Woodhead Publishing Limited, p. 80-116.

Sun Z, Li S, Chen W, Zhang J, Zhang L, Sun W, Wang Z (2021) Plant Dehydrins: Expression, Regulatory Networks, and Protective Roles in Plants Challenged by Abiotic Stress. Int. J. Mol. Sci. 22: 12619.

Syngai G, Barma P, Bharali R, Dey S (2013) BLAST: An introductory tool for students to Bioinformatics Applications BLAST : An introductory tool for students to Bioinformatics Applications. Keanean J. Sci. 2: 67-76.

Teh BT, Lim K, Yong CH, Ng CCY, Rao SR, Rajasegaran V, Lim WK, Ong CK, Chan K, Cheng VKY, Soh PS, Swarup S, Rozen SG, Nagarajan N, Tan P (2017) The draft genome of tropical fruit durian (*Durio zibethinus*). Nat. Genet. 1-9.

Trapnell C, Roberts A, Goff L, Perte G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7: 562-578.

Weirick T, Militello G, Müller R, John D, Dimmeler S, Uchida S (2016) The identification and characterisation of novel transcripts from RNA-seq data. Brief. Bioinform. 17: 678-685.

Yang S, Yang T, Tang Y, Aisimutuola P, Zhang G, Wang B, Wang J, Yu Q (2020) Transcriptomics profile analysis of non-coding RNAs involved in Capsicum chinense Jacq. Fruit ripening. Sci. Horticult. 264.

Yazhini A (2018) Small Open Reading Frames. Resonance 23: 57-67.

Yu Y, Zhang Y, Chen X, Chen Y (2019) Plant Noncoding RNAs: Hidden Players in Development and Stress Responses. Annu. Rev. Cell Dev. Biol. 35: 407-31

Zhao Z, Zang S, Zou W, Pan YB, Yao W, You C, Que Y (2022) Long Non-Coding RNAs: New Players in Plants. Int. J. Mol. Sci. 23: 9301.

Durian Reference Genome- BioProject (PRJNA400310) https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA400310 https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002303985.1/ url:ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/303/985/GCF_002303985.1_Duzib1.0/GCF_002303985.1_Duzib1.0_genomic.fna.gz

Supplementary Materials

Table S1 – Two hundred eighty (280) unknown transcripts from Cuffdiff (XLS)

Table S2 – One hundred eighty-three (183) unknown transcripts that hit for BLASTn (XLS)

Table S3 – Ninety-seven (97) unknown transcripts that have no hits for BLASTn (XLS)

Table S4 – Seventy (70) unknown transcripts that have InterPro Scan results (XLS)
Table S5 – Twenty-seven (27) sequences with no BLASTx hits and no IPS match (XLS
Table S6 – Results analysis of sequences with no BLASTx hit and no IPS match (XLS)
Table S7 – BlastX results and functional annotation of up-regulated unknown transcripts in YSMS
Table S8 – BlastX results and functional annotation of up-regulated unknown transcripts in YSRS
Table S9 – GO analysis of up-regulated unknown transcript in the developmental stage.
Table S10 – GO analysis of up-regulated unknown transcript in the ripening stage.