Scientia Horticulturae 327 (2024) 112846

Contents lists available at ScienceDirect

Scientia Horticulturae

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

Research Paper

An eFP reference gene expression atlas for mangosteen

Ching-Ching Wee [a,b], Asher Pasha [c], Nicholas J. Provart [c], Nor Azlan Nor Muhammad [a], Vijay Kumar Subbiah [b], Masanori Arita [d], Hoe-Han Goh [a,*]

[a] Institute of Systems Biology, Universiti Kebangsaan Malaysia, Bangi, Selangor 43600, Malaysia
[b] Biotechnology Research Institute, Universiti Malaysia Sabah, Kota Kinabalu, Sabah 88400, Malaysia
[c] Department of Cell and Systems Biology/Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Ontario M5S 3B2, Canada
[d] Department of Informatics, National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan

ARTICLE INFO

Keywords:
eFP
Functional genomics
Garcinia
Gene expression
Transcriptome

ABSTRACT

Mangosteen is the queen of fruits with many health benefits and high pharmaceutical potential but remains an underutilized crop. There are several reports of mangosteen organelle genomes and transcriptomes from various tissues but a reference genome or transcriptome is still lacking. In this study, we aimed to generate a publicly accessible reference transcriptome of *Garcinia mangostana* variety Manggis (local fruit in Southeast Asia) to facilitate research in mangosteen functional genomics. *De novo* assembly was performed using Trinity with all the Illumina sequence datasets and generated 369,861 transcripts with an N50 of 1,433 bp. SuperTranscripts and TGICL transcript clustering approaches were taken to reduce redundant transcripts, in which the latter generated a more comprehensive reference transcriptome with a higher N50 value, Benchmarking Universal Single-Copy Orthologs (BUSCO) score, and read mapping rates. A total of 118,165 (43.7 %) unigenes were functionally annotated. To visualize gene expression across different mangosteen tissues, developmental stages, and experiments based on the reference transcriptome, we constructed a mangosteen electronic Fluorescent Pictograph (eFP) browser. This allows users to easily visualize the expression of genes in absolute, relative, and compare modes. In addition, researchers can perform online BLAST search against the in-house BAR SequenceServer for homologous sequences that match mangosteen transcripts to explore corresponding expression patterns via direct links to the eFP browser. This reference transcriptome and eFP browser (accessible at https://bar.utoronto.ca/efp_mangosteen/cgi-bin/efpWeb.cgi) provide a useful online tool for future research and improvement of mangosteen.

1. Introduction

Mangosteen (*Garcinia mangostana* L.) is a climacteric fruit that is well-known as the "Queen of fruits". It has a high market value due to the xanthones in the pericarp possessing pharmaceutical properties (El-Seedi et al., 2010; Ovalle-Magallanes et al., 2017; Shan et al., 2011). Hence, mangosteen becomes the subject of study focusing on the extraction and quantitative analysis of xanthone compounds, their bioactivity (anti-cancer and anti-inflammatory), and bioavailability (Ji et al., 2007; Muchtaridi et al., 2017; Ovalle-Magallanes et al., 2017). Apart from analytical studies of mangosteen extracts, there are emerging omics studies (Jamil et al., 2023, 2021; Mamat et al., 2020). Recently, the complete mitogenome (Wee et al., 2022) and plastome (Wee et al., 2023) of mangosteen have been reported. Several mangosteen transcriptome data were generated from different tissues such as fruits (Abdul-Rahman et al., 2017; Matra et al., 2016, 2019), seeds (Goh et al., 2019; Mazlan et al., 2018), and calli (Mahdavi-Darvari and Noor, 2016). These data are publicly available and accessible from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and DNA Data Bank of Japan (DDBJ) SRA (DRA) database. However, large data analysis requires high-performance computing guided by scientific knowledge for interpretation. Hence, these data are largely inaccessible for many researchers in the field.

The development of an electronic Fluorescent Pictograph (eFP) browser (Winter et al., 2007) helps to tackle this issue. eFP browser is a useful web-based tool for the visualization and interpretation of gene expression data generated from both microarray and RNA sequencing (RNA-seq) datasets of any organisms (Winter et al., 2007). As high-throughput datasets are depicted in pictures of different samples, it is easier for researchers, especially non-bioinformaticians to facilitate

* Corresponding author.
E-mail address: gohhh@ukm.edu.my (H.-H. Goh).

https://doi.org/10.1016/j.scienta.2024.112846
Received 12 October 2023; Received in revised form 26 December 2023; Accepted 2 January 2024

data interpretation and hypothesis generation. As of July 2023, eFP browsers have been developed for twenty-two plants, such as Arabidopsis (Winter et al., 2007), kiwi (Brian et al., 2021), pineapple (Mao et al., 2018), strawberry (Hawkins et al., 2017), and tomato (Fernandez-Pozo et al., 2017). However, an equivalent database is unavailable for mangosteen RNA-seq expression data visualization.

By integrating all the mangosteen transcriptome data from a total of 56 public datasets, we established a mangosteen eFP browser that enables the visualization of gene expression in seed development and germination, fruit ripening, non-embryonic *vs.* somatic embryogenic calli, aril *vs.* rind tissues, and diseased *vs.* normal aril/rind tissues. The mangosteen eFP browser was set up on the Bio-Analytic Resource for Plant Biology (BAR) server (https://bar.utoronto.ca/) and integrated with reference mangosteen sequences in the BAR SequenceServer. Here, we describe the generation of the reference transcriptome and the construction of the mangosteen eFP browser with three example use cases to illustrate various features in utilizing this online tool for functional genomics investigation.

2. Material and methods

2.1. Data collection

Mangosteen RNA samples were sequenced by two research groups: (1) Universiti Kebangsaan Malaysia (UKM), Malaysia and (2) Bogor Agricultural University, Indonesia, using Illumina and Ion Torrent sequencing platforms, respectively. These data were obtained from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) and DNA Data Bank of Japan (DDBJ) SRA (DRA) (Table S1–3). For studies from UKM, the mangosteen trees of variety Manggis were planted from seed plantlet tissue culture and grown at the experimental plot (2°55′09.0″N 101°47′04.8″E) of Universiti Kebangsaan Malaysia, Bangi.

2.2. RNA-seq data processing and analysis

Trimmomatic v0.39 (Bolger et al., 2014) was used to obtain clean Illumina data. It removed the adaptor sequences and low-quality reads with default parameters (TruSeq3-PE-2.fa:2:30:10 SLIDINGWINDOW:4:5 LEADING:5 TRAILING:5 MINLEN:25). Besides, clean Ion Torrent reads were also obtained using Trimmomatic with minor modifications (iontorrent.fa:2:30:10 SLIDINGWINDOW:4:20 LEADING:20 TRAILING:20 MINLEN:25).

Next, the clean Illumina data were used for *de novo* transcriptome assembly with Trinity v.2.9.1 (Grabherr et al., 2011) using default settings and additional parameters (normalize_by_read_set; min_kmer_cov=2; no bowtie) to generate a single Trinity assembly. Next, TGICL (Pertea et al., 2003) and SuperTranscripts (Davidson et al., 2017) were used to obtain non-redundant unigenes using default settings. A total of three approaches were used to assess the quality of the generated reference transcriptomes: (1) contig N50 value, (2) Benchmarking Universal Single-Copy Orthologs (BUSCO) using gVolante ver.1.2.1 (Nishimura et al., 2017), and (3) read mapping rates of both Illumina and Ion Torrent RNA-seq clean reads against the reference transcriptome. BUSCO version 4.1.4 was used to access the completeness of the reference transcriptome generated from TGICL against several databases (Viridiplantae, Embryophyta, Eudicotyledons, and Solanaceae version 10).

For transcript abundance estimation, the raw reads were aligned against the reference transcriptome generated from TGICL using script "Bowtie2" (default parameters) found in the Trinity toolkit v.2.9.1 (Grabherr et al., 2011). Then, gene expression levels in transcript per million (TPM) values were obtained using the salmon tool found in the Trinity toolkit v.2.9.1 (Grabherr et al., 2011).

2.3. Functional annotation

Trinotate v3.2.1 (Bryant et al., 2017) was used to perform functional annotation using the non-redundant genes and transdecoder-predicted proteins. The databases used included: (1) Swiss-Prot (Annotated protein sequence database, 2020_05), (2) GO (Gene Ontology, Release 2020–11–18), and (3) KEGG (Kyoto Encyclopedia of Genes and Genomes). Araport11 CDS and peptide (Release 2019–07–11) were also used as queries to perform local BLASTn and BLASTp (ncbi-blast-2.11.0) using the default command lines (E-value=1e-5). In addition, SignalP and TMHMM were used to predict the protein signal peptide and transmembrane domain, respectively. HMMER v3.0 was used to identify the protein domain from PFAM (Release 2020–05–02). An online KO annotation server (KAAS) (accessed on 28 Dec 2020) based on the single-directional best hit (SBH) method against the plant KEGG GENES data set (*Arabidopsis thaliana*, banana, durian, papaya, and tomato) (Moriya et al., 2007) was used to obtain KEGG Orthology (KO) assignment. Lastly, iTAK v1.7 (Zheng et al., 2016) was used to annotate transcription factor/regulator.

UniProtKB dataset (Release 2022_05) was used to identify xanthone-related proteins by using the keywords "xanthone" and "benzophenone synthase". The xanthone-related proteins masterlist was downloaded and screened to exclude (1) protein sequences before benzoate-CoA ligase (BZL) and (2) proteins not related to xanthone biosynthesis. Additional enzymes involved in the plant xanthone synthesis pathway were identified from a review article (Remali et al., 2022). Their protein sequences and accession numbers were extracted and included in the xanthone-related protein masterlist (Supplementary File 1). Xanthone-related proteins were identified using local BLASTx (E-value=1e-5). Lastly, local BLASTn (E-value=1e-3) was used to identify plastome (Wee et al., 2023) and mitogenome (Wee et al., 2022) sequences from the reference transcriptome.

2.4. Mangosteen eFP browser construction

To construct web-based visualization of gene expression profiles of different mangosteen tissues/developmental stages/experiments, electronic Fluorescent Pictograph (eFP) software version 1.6.0 (Winter et al., 2007) (https://github.com/BioAnalyticResource/eFP) was used. The four input files include: (1) images in both Portable Network Graphics (.png) format, (2) Extensible Markup Language (XML) control file, (3) gene expression database, and (4) gene description file.

GNU Image Manipulation Program (GIMP) image-editing software v2.10.32 (https://www.gimp.org/) was used to draw the diagrams of mangosteen tissues (seeds, fruits, and calli). The diagrams were then converted to PNG format. The transcriptome data with expression value in TPM was prepared in an Excel spreadsheet (Supplementary File 2) and incorporated into the Bio-Analytic Resource for Plant Biology (BAR) in-house server. Each column corresponds to a sample while each row corresponds to one gene. Hence, the first column shows the gene identifiers while the first row contains the sample names. All gene descriptions were prepared in another Excel spreadsheet (Supplementary File 3). An XML-based configuration file was set up for each figure to describe the tissue type, the development stage of a particular tissue, and its unique color code. The mangosteen eFP browser was set up online via the BAR server at https://bar.utoronto.ca/efp_mangosteen/cgi-bin/efpWeb.cgi. Users can explore the expression patterns of sequences that match the mangosteen unigene sequences by performing local BLAST search via the BAR SequenceServer 2.1.0 (https://bar.utoronto.ca/blast) (Priyam et al., 2019).

2.5. Features of mangosteen eFP browser

In the mangosteen eFP browser, three modes can be chosen: (1) "Absolute", (2) "Relative", and (3) "Compare". Individual gene expression in TPM value is directly compared to the highest TPM value

recorded in the sample set of a particular image under the "Absolute" mode. Red indicates a high expression level while yellow indicates a low expression level.

The Log_2 fold-change (FC) of a tissue expression level against the control signal is displayed under the "Relative" mode. Here, the control signal refers to the median of an individual gene of the same sample set. In diseased *vs.* normal eFP Browser, the control signal refers to the mean of the normal aril/rind under control and treatment conditions (FC = mean TPM value / median or mean of control signal).

Lastly, two gene identifiers are used as input under the "Compare" mode. This mode compares the primary gene relative expression levels against the secondary gene relative expression level (FC = primary gene / secondary gene).

3. Results

3.1. De novo transcriptome assembly of mangosteen RNA-seq datasets

There were four mangosteen experiments performed using Illumina sequencing with a total of 20 RNA-seq datasets, including seed development (week (w) 8, 10, 12, and 14 after anthesis), seed germination (day (d) 0, 3, 5, and 7 after sowing), fruit ripening (ripening stage (s) 0, 2, and 6), and calli (non-embryogenic and somatic embryogenic) (Table S1–2). The raw sequence data were trimmed to generate 797,815,733 clean reads (85.8 %) for *de novo* transcriptome assembly (Table S3). A total of 369,861 transcripts with an N50 of 1433 bp were generated by Trinity (Table 1).

To minimize redundant transcripts, SuperTranscripts and TGICL transcript clustering were performed, which generated 209,226 and 270,208 non-redundant unigenes, respectively (Table 1). Meanwhile, the N50 value of unigenes generated by TGICL (1609 bp) was higher than SuperTranscripts (946 bp).

3.2. Reference transcriptome quality assessment

Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis was performed via an online portal for completeness assessment, gVolante, to assess the quality of the transcriptome sequences generated from Trinity, SuperTranscripts, and TGICL. The number of complete sequences based on Embryophyta orthologous database version 9 (odb9) was comparable between Trinity (1316) and TGICL (1318) but higher than SuperTranscripts (1009) (Fig. 1). The latest version 10 orthologous databases (odb10) were used for more in-depth analysis of the TGICL-assembled transcriptome (Table 2), which showed higher number of complete sequences for Embryophyta (93.2 %) than version 9 and over 90 % complete sequences for Eudicotyledons.

On the other hand, we also assessed the transcriptome completeness based on sequencing read alignment. Both Illumina paired-end and unpaired reads showed higher mapping rates against the TGICL reference transcriptome than that of SuperTranscripts (Table S4). As for the Ion Torrent data (Matra et al., 2016, 2019), 423,492,414 clean reads from 760,917,051 raw reads also showed higher mapping rates against TGICL-generated reference than SuperTranscripts (Table S5). Based on the highest N50 value, BUSCO score, and read mapping rates, the TGICL-generated transcriptome with reduced redundancy was chosen as the reference transcriptome for functional annotation and downstream analysis.

Table 1
Summary statistics of assembled transcriptomes.

Attribute	Assembly		
	Trinity	SuperTranscripts	TGICL
Total number of unigenes	209,226	209,226	270,208
Total number of transcripts	369,861	–	–
GC percentage	40.58	40.08	40.52
Statistics based on ALL transcript contigs			
N50 (bp)	1433		
Median length (bp)	429		
Average length (bp)	810		
Total assembled bases	299,487,603		
Statistics based on ONLY LONGEST ISOFORM per GENE			
N50 (bp)	789	946	1609
Median length (bp)	348	364	466
Average length (bp)	591	655	881.5
Total assembled bases	123,645,720	137,008,050	238,182,812
Read mapping rates (%)			
Illumina Paired-end	97.0–98.5	85.1–93.3	96.7–98.3
Unpaired (forward)	95.2–99.0	89.0–94.1	94.8–98.8
Ion torrent	82.6–97.9	79.5–97.2	82.2–97.6

3.3. Characteristics of the non-redundant unigenes and functional annotation

Transdecoder predicted 93,923 peptides from 270,208 TGICL-generated unigenes (Table 3). The unigene lengths range from 180 to 19,297 bp with a mean of 881.5 bp and N50 of 1609 bp. Around 26.3 % of unigenes were more than 1000 bp while the majority were less than 400 bp (43.7 %).

A total of 92,494 (34.2 %) unigenes found hits against the Swiss-Prot protein database based on BLASTx search. Local BLASTn and BLASTp search against Arabidopsis CDS and peptide sequences matched 84,984 (31.5 %) unigenes and 58,843 (21.8 %) peptides, respectively. Meanwhile, 54,231 (20.1 %) and 51,494 (19.1 %) predicted peptide sequences matched the Swiss-Prot and Pfam databases, respectively. There were 41,782 (15.5 %) unigenes with KEGG Orthology (KO) assigned using KAAS and 7396 (2.7 %) predicted peptides annotated against the eggNOG database. A total of 5352 unigenes (2 %) were identified to be transcription factors (TFs) or transcriptional regulators (TRs) using iTAK (Zheng et al., 2016).

The xanthone-related and organelle unigenes accounted for 1587 (0.6 %) and 126 (0.05 %) unigenes, respectively. Among these, local BLASTn results showed that 65 unigenes were related to chloroplast genes while 60 unigenes were related to mitochondrial genes. There was one unigene related to both chloroplast and mitochondrial genes as this region constitutes plastome-derived sequences (Supplementary File 4).

Overall, 118,165 (43.7 %) unigenes were functional annotated. Despite lower than half of all unigenes were annotated, this is the most comprehensive functional annotation done on mangosteen transcriptome from all the available Illumina data to date. Transcript abundance was estimated in transcripts per kilobase million (TPM) values based on the read alignment of both Illumina and Ion Torrent sequencing data. To allow easy public access to these mangosteen gene expression data, an eFP browser was constructed.

3.4. Description of mangosteen eFP browser

For the construction of the mangosteen eFP Browser (Fig. 2), there were six data sources (Fig. 3). The .png image file of the corresponding data source will be displayed when the "Data Source" is chosen. Gene expression can be viewed using three modes: "Absolute", "Relative", and "Compare". The primary gene ID must be entered before the "Go" button is clicked to view the gene expression. Both primary and secondary gene IDs must be entered when the "Compare" mode is chosen. The gene description of the gene ID is given below the image. A direct link to the bioproject/biosample of the image is also provided by clicking the image (Fig. 2).

3.5. Examples of eFP browser use cases

In the "Absolute" mode under the data source "Fruit Ripening", when the Primary Gene ID "DN23451" encoding a phenylalanine ammonia-lyase (*PAL6*) gene was entered, the fruit at S0 was yellow while the fruit at S2 was red (Fig. 4A). This indicates *DN23451* was highly

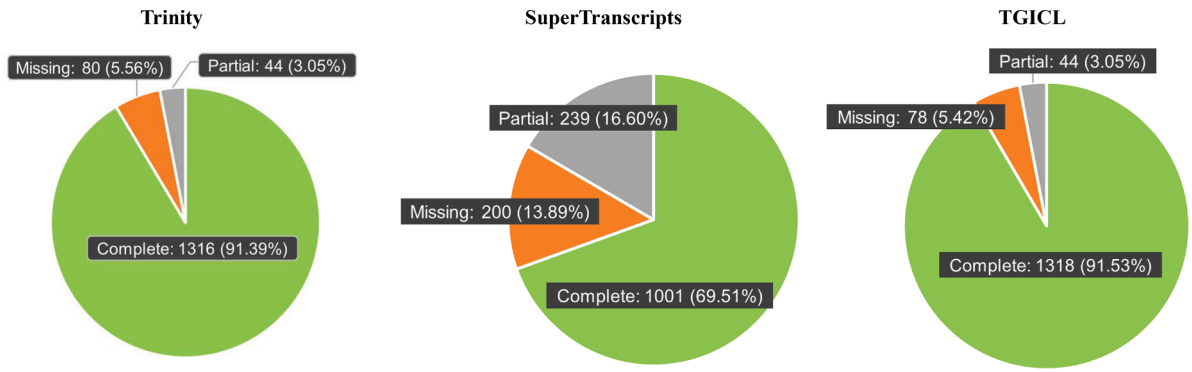


Fig. 1. BUSCO analysis of different transcriptomes using Embryophyta_odb9 database.

Table 2
BUSCO analysis of TGICL-generated trancriptome against several databases.

Database	Complete (C)		Complete and single-copy (S)		Complete and duplicated (D)		Fragmented (F)		Missing (M)		N BUSCO group search
	%	n	%	n	%	n	%	n	%	n	
Viridiplantae_odb10	97.0	412	50.4	214	46.6	198	2.1	9	0.9	4	425
Embryophyta_odb10	93.2	1505	49.6	801	43.6	704	4.1	66	2.7	43	1614
Eudicotyledons_odb10	90.8	2111	49.7	1156	41.1	955	3.1	73	6.1	142	2326

Table 3
Summary statistics of mangosteen *de novo* transcriptome assembly and functional annotation.

Attribute	Number	%
Raw reads	930,153,027	
Clean reads	797,815,733	85.8
Percentage GC		40.5
Number of unigenes	270,208	
Number of predicted peptides	93,923	
Unigene length (bp)		
Total	238,182,812	
Range	180–19,297	
N50	1609	
Mean	881.5	
Length distribution (bp)		
<299	78,950	29.2
300–399	39,077	14.5
400–599	41,932	15.5
600–799	23,625	8.7
800–999	15,507	5.7
1000–1999	40,011	14.8
>2000	31,106	11.5
Functional annotation		
Swiss-Prot (BLASTx)	92,494	34.2
Swiss-Prot (BLASTp)	54,231	20.1
Pfam	51,494	19.1
AGI (BLASTn)	84,984	31.5
AGI (BLASTp)	58,843	21.8
Transcription factor/regulator (iTAK v1.7)	5352	2.0
eggNOG	7396	2.7
KO (KAAS)	41,782	15.5
Xanthone-related unigenes	1587	0.6
Organelle unigenes	126	0.05
Annotated	118,165	43.7
Unannotated	152,043	56.3

expressed in fruit at S2. The expression value can be viewed by clicking the "Table of Expression Values" located below the image (Fig. 4A). The table indicated that *DN23451* was significantly upregulated as the expression (TPM) values at S0 and S2 were 10.83 and 115.87, respectively.

The mean expression levels for samples with replicates and the standard deviation (STDEV) values are shown in the table (Fig. 4B). Besides, a link ("To the Experiment") is given for users to access the bioproject/biosample of the experiment. Additionally, both expression and standard deviation values can be depicted in a chart (Fig. 4C) by clicking "Chart of Expression Values" located below the image (Fig. 4A).

In the second example, a *Pectinesterase 2* (*PME2*) gene was reported as significantly upregulated in translucent flesh disorder (TFD)-aril compared to normal aril (Matra et al., 2019), which is depicted in Fig. 5A. Here, the data source "Diseased *vs.* Normal" was selected and the Primary Gene ID "DN51661" was entered with "Relative" mode selected. TFD-aril in "aril under control condition" was compared against control. The red color indicates a relatively higher expression (Fig. 5A). Log_2FC of 7.24 under group 1 as indicated in the table (Fig. 5B) showed that *PME* was significantly upregulated in diseased aril compared to normal aril under the control condition.

There are four groups of diseased *vs.* normal aril/rind under control and treatment conditions comparison in the "Table of Expression Values" (Fig. 5B). The sample and control signals for the control in each respective group were the same. Diseased tissue is compared against normal tissue that acts as a control in each group. Hence, the sample signal is the average expression value of the particular tissue replicates. The difference in signal (sample signal minus control signal) and fold-change can be visualized by clicking "Chart of Expression Values" (Fig. 5C).

For the third example, when primary gene ID DN23451 and secondary gene ID DN257464 are entered under the "Compare" mode, the relative expression level of the *PAL* gene from the first example *DN23451* (*PAL6*) is compared against that of a leaf developmental *YABBY* gene *DN257464* (*YAB5*) (Fig. 6A). The scale bar indicates the Log_2 ratio of the fold-change (FC) with red indicating high Log_2FC while blue indicates low/negative Log_2FC. *DN23451* unigene showed a lower expression compared to *DN257464* at S0 but a higher expression at S2 (Fig. 6A).

The "Table of Expression Values" shows the Log_2 ratio of FC and FC (Fig. 6B). Log_2 ratio of FC in tissue at S0 was −0.76, 0.44 at S2, and −0.43 at S6, which indicates that the expression of *DN23451* is lower compared to *DN257464* at S0 and S6 but higher at S2. The FC of the unigenes in comparison can be visualized by clicking "Chart of Expression Values" (Fig. 6C). FC values lower than 1 indicates a relatively lower expression of the primary gene than the secondary gene.

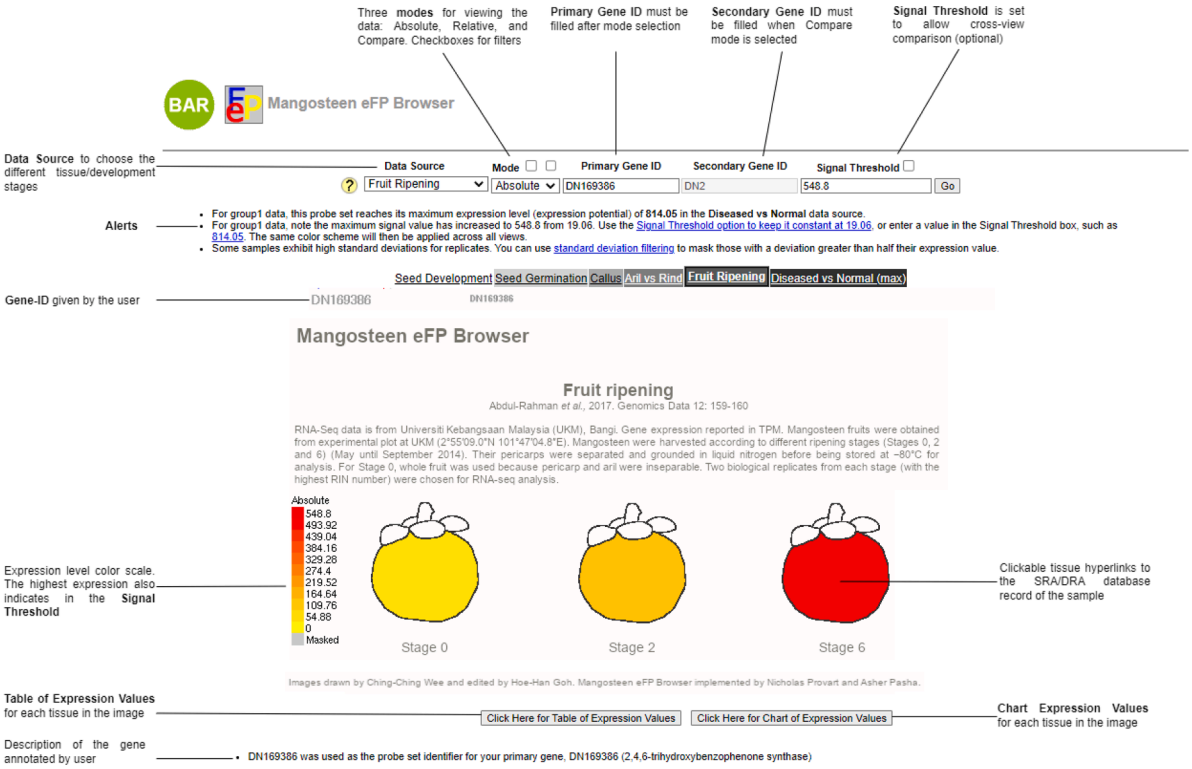


Fig. 2. Mangosteen eFP browser. Different features of eFP browser are annotated.

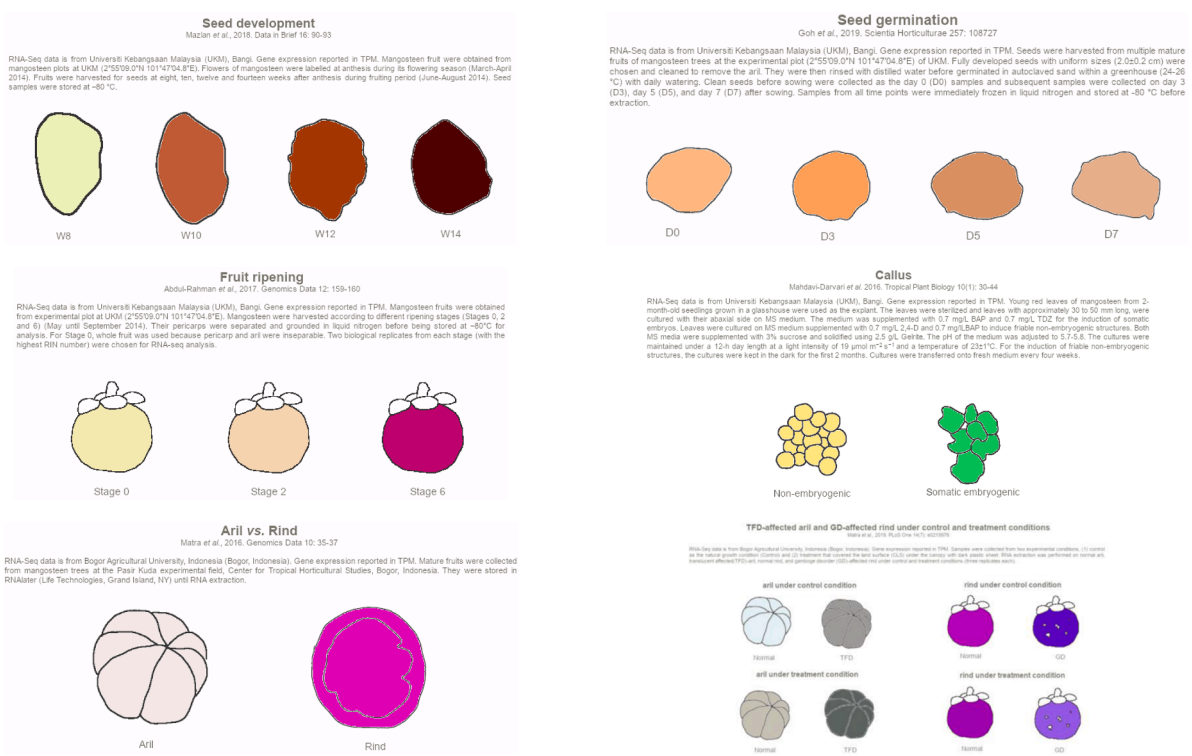
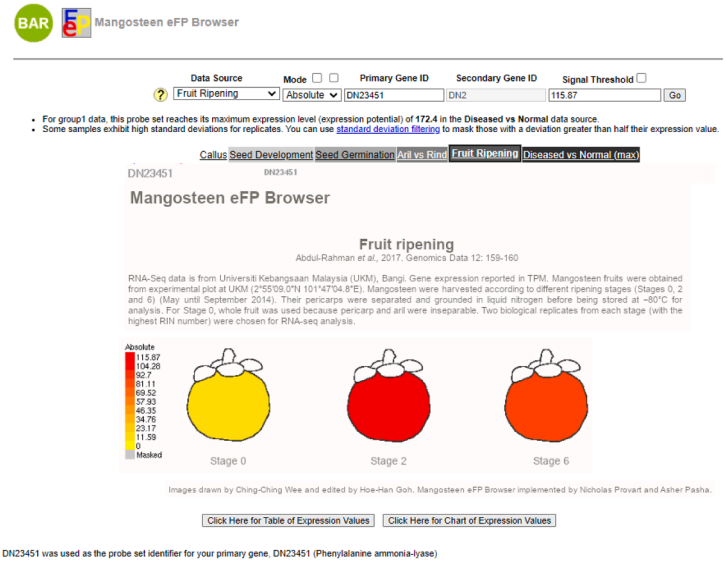


Fig. 3. Images of different tissues/development stages/experiments in the mangosteen eFP browser.

3.6. Local BLAST search against BAR SequenceServer

Apart from the input of targeted unigene ID into the eFP browser, users can also search for homologous gene sequences from the mangosteen reference transcriptome using the BAR SequenceServer (https://bar.utoronto.ca/blast). For example, a partial sequence of unigene *DN1* was used as an input for the query field in BAR SequenceServer (Fig. S1A). The sequence type was automatically detected as a nucleotide sequence with "Nucl" selected under the nucleotide databases. Users can execute the search by clicking the "BLAST". The report shows the

(A)



(B)

Group #	Tissue	Expression Level	Standard Deviation	Samples	Links
1	Fruit ripening stage 0	10.83	10.41	GmS0_1, GmS0_2	To the Experiment
1	Fruit ripening stage 2	115.87	2.16	GmS2_1, GmS2_2	To the Experiment
1	Fruit ripening stage 6	85.18	2.56	GmS6_1, GmS6_2	To the Experiment

(C)

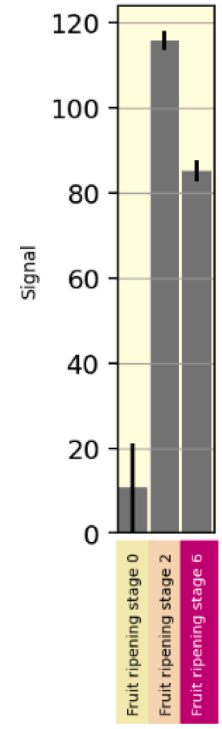
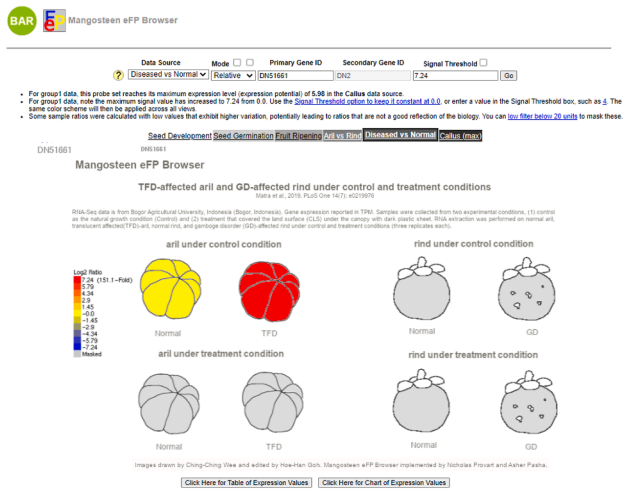


Fig. 4. Expression of *DN23451* (*PAL6*) unigene. (A) Image of *DN23451* (*PAL6*) expression under the "Fruit Ripening" data source. (B) Expression level and standard deviation value in the expression value table. (C) Expression chart with mean expression level (bars) and STDEV (error bars).

(A)



(B)

Group #	Tissue	Sample signal	Control signal	Log2 Ratio	Fold-Change	Samples	Links
1	Normal aril under control condition	0.02	0.02	0.01	1.01	ANC_1, ANC_2, ANC_3	To the Experiment
1	Translucent flesh disorder-affected aril under control condition	3.19	0.02	7.24	151.66	TFDA_AC_1, TFDA_AC_2, TFDA_AC_3	To the Experiment
2	Normal aril under treatment condition	0.0	0.0	nan	nan	ANT_1, ANT_2, ANT_3	To the Experiment
2	Translucent flesh disorder-affected aril under treatment condition	0.46	0.0	nan	nan	TFDA_AT_1, TFDA_AT_2, TFDA_AT_3	To the Experiment
3	Normal rind under control condition	0.0	0.0	nan	nan	RNC_1, RNC_2, RNC_3	To the Experiment
3	Gamboge disorder-affected rind under control condition	0.0	0.0	nan	nan	GDA_RC_1, GDA_RC_2, GDA_RC_3	To the Experiment
4	Normal rind under treatment condition	0.0	0.0	nan	nan	RNT_1, RNT_2, RNT_3	To the Experiment
4	Gamboge disorder-affected rind under treatment condition	0.00	0.0	nan	nan	GDA_RT_1, GDA_RT_2, GDA_RT_3	To the Experiment

(C)

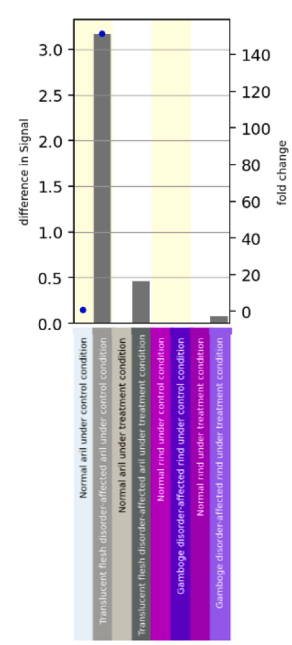
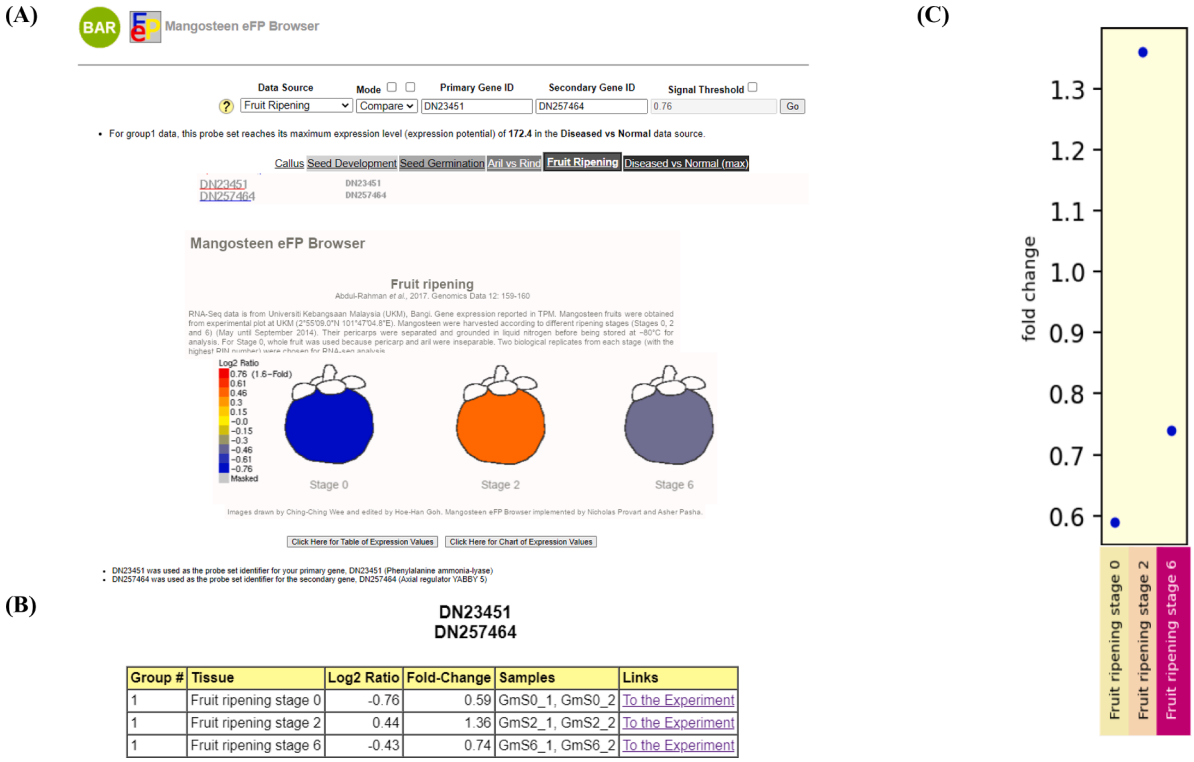


Fig. 5. Expression of *DN51661* (*PME2*). (A) Image of *DN51661 (PME2)* expression under the "Diseased *vs*. normal" data source. (B) Sample signal, control signal, Log_2 ratio, and fold-change in the expression value table. (C) Expression chart with sample signal (bar) and fold-change (blue dot).

summary of the query with database and parameters (default settings), BLAST result, and alignment result (Fig. S1B). *DN1* was correctly detected as the top hit with *DN4692* identified as the second-best hit. The expression of this unigene in different mangosteen samples can be explored by clicking the "eFP" link, which can also be shared via a hyperlink.

4. Discussion

There are a total of 56 mangosteen transcriptome data sets (Table S1–3) that can be found in the public database. They were generated by two groups of researchers (Universiti Kebangsaan Malaysia, Malaysia and Bogor Agricultural University, Indonesia). These



Group #	Tissue	Log2 Ratio	Fold-Change	Samples	Links
1	Fruit ripening stage 0	-0.76	0.59	GmS0_1, GmS0_2	To the Experiment
1	Fruit ripening stage 2	0.44	1.36	GmS2_1, GmS2_2	To the Experiment
1	Fruit ripening stage 6	-0.43	0.74	GmS6_1, GmS6_2	To the Experiment

Fig. 6. Comparison of gene expression between *DN23451* (*PAL6*) and *DN257464* (*YAB5*). (A) Image of gene comparison under the "Fruit Ripening" data source. (B) Log_2 ratio and fold-change in the expression value table. (C) Expression chart with fold-change (blue dot).

datasets include (1) seed development, (2) seed germination, (3) fruit ripening, (4) calli, (5) aril *vs.* rind, and (6) diseased *vs.* normal fruits (Table S2, Fig. 3).

To analyze gene expression in different samples, *de novo* transcriptome assembly was performed followed by transcriptome quality assessment. From the read mapping results, gene expression levels were estimated based on transcript abundance in TPM values. These tedious steps of data analysis require extensive computing with bioinformatic software and knowledgeable manpower. Hence, these valuable data might not be accessible for researchers without these facilities and knowledge. Therefore, it is imperative to generate a public tool to compile and visualize gene expression for easy interpretation. The development of a user-friendly online eFP browser (Winter et al., 2007) by the University of Toronto provides such a solution.

Mangosteen eFP Browser (Fig. 2) is a useful tool for easy visualization of gene expression. Users can interpret the gene expression results easily and study the gene expression in different tissues. In addition, it allows cross-disciplinary collaboration between researchers, biologists, and farmers (Hawkins et al., 2017). Besides, users can also compare the expression level across different tissues for future investigation and hypothesis generation (Winter et al., 2007).

In this case study, three modes were used to show the gene expression from different data sources. In mangosteen, fruit ripening is accompanied by increased accumulation of anthocyanin (purple-color rind). The biosynthesis of anthocyanin (Zhao and Tao, 2015) involved the phenylalanine-dependent pathway and *PAL* enzyme is responsible for the conversion of phenylalanine into cinnamic acid. The upregulation of *PAL* gene had been reported in transcriptomics analysis (Jamil et al., 2023) and displayed clearly in the mangosteen fruit ripening eFP Browser (Fig. 4).

In the second case study, *PME2* a cell wall-modifying gene was identified as the top ten upregulated gene by comparing TFD-aril *vs.* normal aril (Matra et al., 2019). One of the features of translucent aril is that its aril is firmer than normal and contains higher sodium carbonate (Na_2CO_3) soluble pectin than normal aril (Dangcham and Siripanich, 2000). This coincided with the significant upregulation of *PME2* (Matra et al., 2019) (Fig. 5).

The third use case explored the relative expression between *DN23451* (*PAL6*) and *DN257464* (*YAB5*) unigenes in mangosteen. *Ac*YAB5 has been reported to be involved in star fruit development (Li et al., 2022). The levels of anthocyanin and total flavonoid were regulated by *Aa*YABBY5 in *Artemisia annua* (Kayani et al., 2021). This might explain why the Log_2 ratio of *PAL6* relative expression value against *YAB5* was lower at fruit ripening stages S0 and S6 (Fig. 6). Meanwhile, *PAL6* expression was higher than *YAB5* at S2, which coincides with the early fruit ripening stage. On the other hand, gene expression of any matching query sequences in BLAST searches against the BAR SequenceServer can be explored via the results linked to the eFP browser of all available species in the database for comparative analysis. This demonstrates the usefulness of eFP browser in exploring homologous gene functions for further functional validation in mangosteen, such as the involvement of *YAB5* in mangosteen fruit development.

Conclusion

In this study, we have generated a reference transcriptome from TGICL clustering of Trinity *de novo* assembly of all publicly available mangosteen Illumina datasets. The expression levels of all unigenes from different experiments were profiled based on TPM values and incorporated into the mangosteen eFP Browser for displaying gene expression data in pictographs with different modes. The mangosteen reference transcriptome in the BAR SequenceServer allows researchers to perform custom local BLAST search of homologous sequences to explore the corresponding gene expression in multiple mangosteen tissues and across different available species in the database. This useful online tool will facilitate future functional genomics investigations of mangosteen for crop improvement.

Data availability

The Supplementary Files and sequences of the reference transcriptome and predicted peptide are available in figshare (10.6084/m9.figshare.23161250).

CRediT authorship contribution statement

Ching-Ching Wee: Writing – review & editing, Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Asher Pasha:** Writing – review & editing, Visualization, Resources. **Nicholas J. Provart:** Writing – review & editing, Visualization, Software, Resources. **Nor Azlan Nor Muhammad:** Writing – review & editing, Supervision. **Vijay Kumar Subbiah:** Writing – review & editing, Supervision. **Masanori Arita:** Writing – review & editing, Supervision. **Hoe-Han Goh:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Hoe-Han Goh reports financial support was provided by National University of Malaysia.

Data availability

All data are publicly accessible.

Acknowledgments

We would like to acknowledge the support of this research by Universiti Kebangsaan Malaysia (UKM) Research University grant DIP-2020-005 (H—HG). In addition, we also would like to acknowledge the hosting of the Mangosteen eFP Browser at the University of Toronto (AP and NJP).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.scienta.2024.112846.

References

Abdul-Rahman, A., Goh, H.H., Loke, K.K., Mohd Noor, N., Aizat, W.M., 2017. RNA-seq analysis of mangosteen (*Garcinia mangostana* L.) fruit ripening. Genom Data 12, 159–160.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Brian, L., Warren, B., McAtee, P., Rodrigues, J., Nieuwenhuizen, N., Pasha, A., David, K. M., Richardson, A., Provart, N.J., Allan, A.C., 2021. A gene expression atlas for kiwifruit (*Actinidia chinensis*) and network analysis of transcription factors. BMC Plant Biol. 21, 1–11.

Bryant, D.M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M.B., Payzin-Dogru, D., Lee, T.J., Leigh, N.D., Kuo, T.H., Davis, F.G., Bateman, J., Bryant, S., Guzikowski, A. R., Tsai, S.L., Coyne, S., Ye, W.W., Freeman Jr., R.M., Peshkin, L., Tabin, C.J., Regev, A., Haas, B.J., Whited, J.L., 2017. A tissue-mapped axolotl *de novo* transcriptome enables identification of limb regeneration factors. Cell. Rep. 18, 762–776.

Dangcham, S., Siripanich, J., 2000. Mechanism of Flesh Translucent Disorder Development of Mangosteen Fruit. KU Annual Conf. Proc.

Davidson, N.M., Hawkins, A.D.K., Oshlack, A., 2017. SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. Genome Biol. 18, 148.

El-Seedi, H.R., El-Barbary, M.A., El-Ghorab, D.M.H., Bohlin, L., Anna-Karin, B.-K., Goransson, U., Verpoorte, R., 2010. Recent insights into the biosynthesis and biological activities of natural xanthones. Curr. Med. Chem. 17, 854–901.

Fernandez-Pozo, N., Zheng, Y., Snyder, S.I., Nicolas, P., Shinozaki, Y., Fei, Z., Catala, C., Giovannoni, J.J., Rose, J.K., Mueller, L.A., 2017. The tomato expression atlas. Bioinformatics 33, 2397–2398.

Goh, H.-H., Abu Bakar, S., Kamal Azlan, N.D., Zainal, Z., Mohd Noor, N., 2019. Transcriptional reprogramming during *Garcinia*-type recalcitrant seed germination of *Garcinia mangostana*. Sci. Hortic. 257, 108727.

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29, 644–652.

Hawkins, C., Caruana, J., Li, J., Zawora, C., Darwish, O., Wu, J., Alkharouf, N., Liu, Z., 2017. An eFP browser for visualizing strawberry fruit and flower transcriptomes. Hortic. Res. 4, 17029.

Jamil, I.N., Abdul-Rahman, A., Goh, H.-H., Aizat, W.M., 2023. Transcriptomics analysis of mangosteen ripening revealed active regulation of ethylene, anthocyanin and xanthone biosynthetic genes. Postharvest Biol. Technol. 198, 112257.

Jamil, I.N., Sanusi, S., Mackeen, M.M., Noor, N.M., Aizat, W.M., 2021. SWATH-MS proteomics and postharvest analyses of mangosteen ripening revealed intricate regulation of carbohydrate metabolism and secondary metabolite biosynthesis. Postharvest Biol. Technol. 176, 111493.

Ji, X., Avula, B., Khan, I.A., 2007. Quantitative and qualitative determination of six xanthones in *Garcinia mangostana* L. by LC–PDA and LC–ESI-MS. J. Pharm. Biomed. Anal. 43, 1270–1276.

Kayani, S.I., Shen, Q., Rahman, S.U., Fu, X., Li, Y., Wang, C., Hassani, D., Tang, K., 2021. Transcriptional regulation of flavonoid biosynthesis in *Artemisia annua* by *Aa*YABBY5. Hortic. Res. 8, 257.

Li, C., Dong, N., Shen, L., Lu, M., Zhai, J., Zhao, Y., Chen, L., Wan, Z., Liu, Z., Ren, H., 2022. Genome-wide identification and expression profile of YABBY genes in Averrhoa carambola. PeerJ 10, e12558.

Mahdavi-Darvari, F., Noor, N.M., 2016. New insight into early somatic embryogenesis of mangosteen (*Garcinia mangostana*) through *de novo* and comparative transcriptome analyses. Trop. Plant Biol. 10, 30–44.

Mamat, S.F., Azizan, K.A., Baharum, S.N., Noor, N.M., Aizat, W.M., 2020. GC–MS and LC-MS analyses reveal the distribution of primary and secondary metabolites in mangosteen (*Garcinia mangostana* Linn.) fruit during ripening. Sci. Hortic. 262, 109004.

Mao, Q., Chen, C., Xie, T., Luan, A., Liu, C., He, Y., 2018. Comprehensive tissue-specific transcriptome profiling of pineapple (*Ananas comosus*) and building an eFP-browser for further study. PeerJ 6, e6028.

Matra, D.D., Kozaki, T., Ishii, K., Poerwanto, R., Inoue, E., 2016. *De novo* transcriptome assembly of mangosteen (*Garcinia mangostana* L.) fruit. Genom. Data 10, 35–37.

Matra, D.D., Kozaki, T., Ishii, K., Poerwanto, R., Inoue, E., 2019. Comparative transcriptome analysis of translucent flesh disorder in mangosteen (*Garcinia mangostana* L.) fruits in response to different water regimes. PLoS ONE 14, e0219976.

Mazlan, O., Abdul-Rahman, A., Goh, H.H., Aizat, W.M., Mohd Noor, N., 2018. Data on RNA-seq analysis of *Garcinia mangostana* L. seed development. Data Brief 16, 90–93.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35, W182–W185.

Muchtaridi, M., Puteri, N.A., Milanda, T., Musfiroh, I., 2017. Validation analysis methods of α-mangostin, γ-mangostin and gartanin mixture in mangosteen (*Garcinia mangostana* L.) fruit rind extract from West Java with HPLC. J. Appl. Pharmaceut. Sci. 7, 125–130.

Nishimura, O., Hara, Y., Kuraku, S., 2017. gVolante for standardizing completeness assessment of genome and transcriptome assemblies. Bioinformatics 33, 3635–3637.

Ovalle-Magallanes, B., Eugenio-Pérez, D., Pedraza-Chaverri, J., 2017. Medicinal properties of mangosteen (*Garcinia mangostana* L.): a comprehensive update. Food Chem. Toxicol. 109, 102–122.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., Quackenbush, J., 2003. TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 19, 651–652.

Priyam, A., Woodcroft, B.J., Rai, V., Moghul, I., Munagala, A., Ter, F., Chowdhary, H., Pieniak, I., Maynard, L.J., Gibbins, M.A., 2019. Sequenceserver: a modern graphical user interface for custom BLAST databases. Mol. Biol. Evol. 36, 2922–2924.

Remali, J., Sahidin, I., Aizat, W.M., 2022. Xanthone biosynthetic pathway in plants: a review. Front. Plant Sci. 13, 916.

Shan, T., Ma, Q., Guo, K., Liu, J., Li, W., Wang, F., Wu, E., 2011. Xanthones from mangosteen extracts as natural chemopreventive agents: potential anticancer drugs. Curr. Mol. Med. 11, 666–677.

Wee, C.-C., Nor Muhammad, N.A., Subbiah, V.K., Arita, M., Nakamura, Y., Goh, H.-H., 2022. Mitochondrial genome of *Garcinia mangostana* L. variety Mesta. Sci Rep 12, 9480.

Wee, C.-C., Nor Muhammad, N.A., Subbiah, V.K., Arita, M., Nakamura, Y., Goh, H.-H., 2023. Plastomes of *Garcinia mangostana* L. and comparative analysis with other *Garcinia* Species. Plants 930.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G.V., Provart, N.J., 2007. An "Electronic Fluorescent Pictograph" browser for exploring and analyzing large-scale biological data sets. PLoS ONE 2, e718.

Zhao, D., Tao, J., 2015. Recent advances on the development and regulation of flower color in ornamental plants. Front. Plant Sci. 6, 261.

Zheng, Y., Jiao, C., Sun, H., Rosli, Hernan G., Pombo, Marina A., Zhang, P., Banf, M., Dai, X., Martin, Gregory B., Giovannoni, James J., Zhao, Patrick X., Rhee, Seung Y., Fei, Z., 2016. iTAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. Mol. Plant 9, 1667–1670.