Genomics Data 10 (2016) 35–37

Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

De novo transcriptome assembly of mangosteen (*Garcinia mangostana* L.) fruit

Deden Derajat Matra [a,b,c], Toshinori Kozaki [d], Kazuo Ishii [d], Roedhy Poerwanto [b], Eiichi Inoue [a,c,*]

[a] United Graduate School of Agricultural Sciences, Tokyo University of Agriculture and Technology, Saiwai, Fuchu, Tokyo 183-8509, Japan
[b] Department of Agronomy and Horticulture, Faculty of Agriculture, Bogor Agricultural University, IPB Dramaga Campus, Bogor, West Java 16680, Indonesia
[c] College of Agriculture, Ibaraki University, 3-21-1 Chuo, Ami, Inashiki, Ibaraki 300-0393, Japan
[d] Department of Applied Biological Science, Faculty of Agriculture, University of Agriculture and Technology, Saiwai, Fuchu, Tokyo 183-8509, Japan

ARTICLE INFO

Article history:
Received 1 September 2016
Accepted 7 September 2016
Available online 9 September 2016

Keywords:
Mangosteen
Fruit
Transcriptome
RNA-Seq
Ion Proton

ABSTRACT

Garcinia mangostana L. (Mangosteen), of the family Clusiaceae, is one of the economically important tropical fruits in Indonesia. In the present study, we performed *de novo* transcriptomic analysis of *Garcinia mangostana* L. through RNA-Seq technology. We obtained the raw data from 12 libraries through Ion Proton System. Clean reads of 191,735,809 were obtained from 307,634,890 raw reads. The raw data obtained in this study can be accessible in DDBJ database with accession number of DRA005014 with bioproject accession number of PRJDB5091. We obtained 268,851 transcripts as well as 155,850 unigenes, having N50 value of 555 and 433 bp, respectively. Transcript/unigene length ranged from 201 to 5916 bp. The unigenes were annotated with two main databases from NCBI and UniProtKB, respectively having annotated-sequences of 73,287 and 73,107, respectively. These transcriptomic data will be beneficial for studying transcriptome of *Garcinia mangostana* L.

Specifications	
Organism/cell line/tissue	*Garcinia mangostana* L./rind and aril
Sex	*NA*
Sequencer or array type	Ion Proton System™
Data format	*Raw data*: *FASTQ file*
Experimental factors	*De novo transcriptome assembly of two tissues from Garcinia mangostana* L.
Experimental features	*Aril and Rind of Garcinia mangostana* L. *fruit were collected for RNA isolation, de novo transcriptome assembly and annotations*
Consent	*N/A*
Sample source location	*Bogor, West Java, Indonesia, S6° 36′ 30.6″ E106° 47′ 03.0″*

1. Direct link to deposited data

The dataset of RNA-seq was deposited at DDBJ Sequence Read Archive (DRA) under with the accession number of DRA005014, and can be accessed at http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA005014.

2. Introduction

Of the tropical fruits found in Indonesia, mangosteen (*Garcinia mangostana* L.) is the most marketable because of its unique taste and flavor. It is also known as the queen of tropical fruits. The edible portion is the white juicy aril, which has a sweet acid taste and is consumed fresh. There is an abundance of polyphenolic compounds in the purple rind of the fruit. Obolskiy et al. [8] assessed the phytochemistry and pharmacological activity of mangosteen fruits and revealed its role as a dietary antioxidant in humans due to the presence of highly polyphenolic compounds such as xanthone. The availability of genomic and transcriptomic data for mangosteen in public databases such as that of the NCBI is limited. However, these are required to better understand the molecular mechanisms and pathways involved in the biosynthesis of beneficial compounds such as polyphenols. Here, we conducted Ion Proton systems-based RNA sequencing analysis to generate the first *de novo*-assembled transcriptome of mangosteen. These transcriptomic data provide useful information to reveal putative genes involved in the biosynthesis of functional compounds such as xanthone and help identify novel genes found in mangosteen.

* Corresponding author at: College of Agriculture, Ibaraki University, 3-21-1 Chuo, Ami, Inashiki, Ibaraki 300-0393, Japan.
E-mail address: eiichi.inoue.a@vc.ibaraki.ac.jp (E. Inoue).

http://dx.doi.org/10.1016/j.gdata.2016.09.003

3. Experimental design, materials and methods

3.1. Plant material

Samples were collected from mature mangosteen trees at the Pasir Kuda experimental field, Center for Tropical Horticultural Studies, Bogor Agricultural University (Bogor, Indonesia) and stored in RNAlater (Life Technologies, Grand Island, NY) until RNA extraction. Samples consisted of two types of tissue from the aril and rind, respectively.

3.2. RNA isolation, library construction, and RNA sequencing

Total RNA was extracted using the hot-borate method [12]. RNA quantity and quality were assessed using the 2100 Bioanalyzer (Agilent, Santa Clara, *CA*). Intact polyadenylated (PolyA) mRNA was subsequently isolated from total RNA using the Dynabeads® mRNA DIRECT™ Micro Purification Kit (Life Technologies, Grand Island, NY). Sequencing libraries from the isolated mRNA were constructed using the Ion Total RNA-Seq Kit v2 and Ion PI™ Template OT2 200 Kit v3 (Life Technologies, Grand Island, NY) following the manufacturer's protocol. RNA-Seq was performed using the Ion PI™ Sequencing 200 kit v3 on the Ion Proton ™ System (Life Technologies, Grand Island, NY).

3.3. Data pre-processing, assembly, and annotation

Quality control for the raw reads was performed using FastQC [1], and high quality reads were obtained using Cutadapt [7] for adaptor trimming and PRINSEQ lite v0.20.4 with specific parameters [11]. These were (1) trimming reads both of 5′- and 3′-ends with quality less than 20 (2) trimming poly-A/T tails having more than 5 bp (3) filtering reads with average quality less than 20 (4) filtering reads having low complexity by Entropy method with threshold at 70 and (5) filtering reads by length ranging from 30 to 230 bp. Ribosomal DNA contaminants were discarded using SortMeRNA v2.1 for the included eight databases [5]. *De novo* assembly of transcriptome with clean reads was carried out using Trinity 2.2.0 with default parameters [3], and *in silico* normalization with 30× coverage was performed.

The longest transcripts were regarded as unigenes for functional annotation. All unigenes were aligned using BLAST+ (version 2.3.0; [2]) against the NCBI non-redundant (nr) databases for both nucleotides and proteins (subset to Viridiplantae, downloaded on May 17, 2016) and the Viridiplantae UniProtKB databases (Swiss-Prot and TrEMBL, downloaded on May 12, 2016) with an E-value cutoff of 10^{-5}. The candidate protein in the coding sequences was extracted using TransDecoder v3.0.0 and the sequences were functionally annotated using Trinotate v3.0 as part of the Trinity functional annotation protocol [4]. Briefly, the protein candidate was searched against several databases: Swiss-Prot to identify known protein with an E-value cut-off of 10^{-5}, PFAM domain database to identify protein domain [10], SignalP to predict signal peptides [9], and tmHMM to predict transmembrane regions [6].

All statistics of reads and assembled sequence were determined (Table 1). Clean reads were thus obtained 191,735,809 (62.3%) from 307,634,890 raw reads. We obtained 268,851 transcripts as well as 155,850 unigenes with N50 lengths of 555 and 433 bp, respectively. Transcript/unigene length ranged from 201 to 5916 bp. All functional annotations are presented in Table 2. The unigenes were annotated with two main databases from NCBI and UniProtKB. Unigenes annotation for nr protein sequences was 73,287 and for nr nucleotide sequences was 39,944. We predicted a total of 35,498 proteins. The BLAST results revealed that 21,316 proteins were annotated by UniProtKB (Swiss-Prot).

Table 2
Functional annotation of mangosteen (*Garcinia mangostana* L.) unigenes.

Database source	Number of sequence (percentage)
Unigenes	
Sequence Number	155,850
- Non-redundant protein (nr) NCBI	73,287 (47.02%)
- Non-redundant Nucleotide (nt) NCBI	39,944 (25.63%)
- Swiss-Prot UniProt	50,330 (32.29%)
- TrEMBL UniProt	73,107 (46.91%)
like-protein sequences	
Sequence Number	35,498
- Swiss-Prot	21,316 (60.04%)
- Pfam	20,169 (56.82%)
- tmHMM	4983 (14.04%)
- signalP	1175 (3.31%)
- RNAMMER	0 (0%)

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

This part of research was carried out with the support of the "Human Resource Development Program in Agricultural Genome Sciences" at the Tokyo University of Agriculture and Technology. We would like to thank to Dr. Darda Efendi as Director of Center for Tropical Horticultural Studies, Bogor Agricultural University for permission to conduct the research. DDM was supported by a MEXT (Ministry of Education, Culture, Sports, Science, and Technology) Japanese Government scholarship during the performance of the research.

Table 1
Read and assembly statistics of rind and aril fruits transcriptome of mangosteen (*Garcinia mangostana* L.)

Features	Rind	Aril
Raw reads (Bases)	176,794,370 (16,861,184,699 bp)	127,840,520 (12,774,960,563 bp)
Trimming and filtering by specific parameter (Bases)	148,514,856 (12,904,647,537 bp)	107,492,143 (9,741,564,258 bp)
Filtering rDNA (Bases)	117,880,842 (10,251,417,048 bp)	73,854,967 (6,655,975,364 bp)
Read length (bp)	35–230	
Total processed reads and bases	191,735,809 (16,907,392,412 bp)	
Number and bases total (bp) of transcripts	268,851/128,161,609	
Number and bases total (bp) of unigenes	155,850/65,042,315	
Length range, average (bp), and N50 (bp) of transcripts	201–5916/476.70/555	
Length range, average (bp), and N50 (bp) of unigenes	201–5916/417.34/433	

References

[1] S. Andrews, FastQC a quality-control tool for high-throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/2010.

[2] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BLAST+: architecture and applications. BMC Bioinformatics 10 (2008) 421.

[3] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-seq data without a reference genome. Nat. Biotechnol. 29 (7) (2011) 644–652.

[4] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, M.D. Macmanes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C.N. Dewey, R. Henschel, R.D. Leduc, N. Friedman, A. Regev, De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8 (8) (2013) 1494–1512.

[5] E. Kopylova, L. Noé, H. Touzet, SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28 (24) (2012) 3211–3217.

[6] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. 305 (3) (2001) 567–580.

[7] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17 (1) (2011) 10–12.

[8] D. Obolskiy, I. Pischel, N. Siriwatanametanon, M. Heinrich, *Garcinia mangostana* L.: a phytochemical and pharmacological review. Phytother. Res. 23 (2009) 1047–1065.

[9] T.N. Petersen, S. Brunak, G.V. Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8 (2011) 785–786.

[10] P.C. Punta, R.Y. Coggill, J. Eberhardt, J. Mistry, C. Tate, N. Boursnell, K.F. Pang, J. Ceric, A. Clements, L. Heger, E.L.L. Holm, S.R. Sonnhammer, A. Eddy, R.D.F. Bateman, The Pfam protein families database Nucleic Acids Research. Database Issue 40 (2012) D290–D301.

[11] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets. Bioinformatics 27 (2011) 863–864.

[12] C.Y. Wan, T.A. Wilkins, A modified hot borate method significantly enhances the yield of high quality RNA from cotton (*Gossypium hirsutum* L.). Anal. Biochem. 223 (1994) 7–12.