Genomics Data 7 (2016) 62–63

Contents lists available at ScienceDirect

Genomics Data

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

DNA-seq analysis of *Garcinia mangostana*

Syuhaidah Abu Bakar, Sureshkumar Sampathrajan, Kok-Keong Loke, Hoe-Han Goh *, Normah Mohd Noor

Institute of Systems Biology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

ARTICLE INFO

Article history:
Received 11 November 2015
Accepted 22 November 2015
Available online 23 November 2015

Keywords:
Apomictic
Genome sequencing
Genome survey
Illumina sequencing
Mangosteen

ABSTRACT

Mangosteen (*Garcinia mangostana* Linn.) is a tropical tree mainly found in South East Asia and considered as "the queen of fruits". The asexually produced fruit is dark purple or reddish in color, with white flesh which is slightly acidic with sweet flavor and a pleasant aroma. The purple pericarp tissue is rich in xanthones which are useful for medical purposes. We performed the first genome sequencing of this commercially important fruit tree to study its genome composition and attempted draft genome assembly. Raw reads of the DNA sequencing project have been deposited to SRA database with the accession number SRX1426419.

Specification	
Subject area	Biology, plant molecular biology
Type of data	Genomic DNA sequences
Organism/Cell line/tissue	*Garcina mangostana* (leaf)
Sequencer type	Illumina HiSeq™ 2000
Data format	Raw sequences (Fastq)
Experimental factors	Experimental plot
Experimental features	DNA-seq dataset for mangosteen genome survey
Sample source location	Malaysia
Data accessibility	SRA database accession SRX1426419 (http://www.ncbi.nlm.nih.gov/sra/SRX1426419)

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/sra/SRX1426419.

2. Value of the data

- *Garcinia mangostana* plant is lacking in molecular genetics information, which hinders genetic studies and crop improvement of this commercially important fruit tree.
- The sequence data are important for genome survey and provide sequence information on the GC content, heterozygosity and estimated genome size.

3. Data

Genome sequences of *G. mangostana* were generated from DNA extract of young leaf tissues. The short reads were filtered, processed, assembled and analyzed as describe in the next section. Raw data for this project were deposited at SRA database with the accession number SRX1426419 (http://www.ncbi.nlm.nih.gov/sra/SRX1426419).

4. Experimental design, materials and methods

4.1. Plant materials

Mangosteen plants were grown under shady environment in experimental plot (2°55′09.0″N 101°47′04.8″E) at Universiti Kebangsaan Malaysia, Bangi. Red young leaf tissues from 4 to 5 months old plant were collected and frozen in liquid nitrogen before stored at −80 °C for DNA extraction.

4.2. DNA extraction and quality control, library preparation and DNA-seq

DNA from leaf samples were extracted using DNeasy Plant mini kit (QIAGEN) based on manufacturer's protocol. Quantity and quality of extracted total DNA were determined using NanoDrop 1000 (Thermo Fisher Scientific Inc., USA) and Agilent 2100 bioanalyzer (Agilent Technologies, USA), respectively.

Paired end reads of 101 bp was generated through the Illumina HiSeq 2000 sequencing platform using the standard DNA library preparation protocol implemented by BGI-Shenzhen, China.

---

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

http://dx.doi.org/10.1016/j.gdata.2015.11.018

Table 1
Statistics of *Garcinia mangostana* sequencing and assembly.

Attributes	Value
Raw reads	
Total number	505,856,290
Total bases (bp)	51,091,485,290
Filtered reads	
Total number	418,812,062
Total bases (bp)	42,300,018,262
N (%)	0.0089
GC (%)	38.14
Q20 (%)	99.19
Q30 (%)	95.43
Minia assembly	
K-mer	41
Number of contigs	281,494
Contig Size	272,873,894
N50 (bp)	1006
Size range (bp)	83–14,015
SSPACE scaffolding	
Number of scaffolds	284,879
Scaffold Size	279,483,966
N50 (bp)	1022

4.3. Raw reads processing and assembly

Raw reads were filtered to remove adapter sequences with sequence pre-processing tool, Trimmomatic [1]. High quality Illumina raw reads with phred score ≥25 were kept for assembly. We have predicted best k-mer length (41 bp) for assembly using KmerGenie [2] and SGA Preqc [3]. *De novo* assembly was done by using Minia assembler v2.0.3 [4] followed by scaffolding using SSPACE [5]. Assembled genome draft was evaluated using CEGMA pipeline [6] by mapping towards core eukaryotic genes in clusters of eukaryotic orthologous groups (KOG) [7]. Table 1 shows the sequencing and assembly statistics.

Conflict of interest

All the authors have approved submission and there are no conflicts of interest.

Acknowledgements

This research was supported by Universiti Kebangsaan Malaysia (UKM) Research University grant AP-2012-018.

References

[1] A.M. Bolger, M. Lohse, B. Usadel, Bioinformatics 30 (2014) 2114–2120.
[2] R. Chikhi, P. Medvedev, Bioinformatics 30 (2014) 31–37.
[3] J.T. Simpson, Bioinformatics 30 (2014) 1228–1235.
[4] R. Chikhi, G. Rizk, Algorithms Mol. Biol. 8 (2013) 22.
[5] M. Boetzer, C.V. Henkel, H.J. Jansen, D. Butler, W. Pirovano, Bioinformatics 27 (2011) 578–579.
[6] G. Parra, K. Bradnam, I. Korf, Bioinformatics 23 (2007) 1061–1067.
[7] E.V. Koonin, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, D.M. Krylov, K.S. Makarova, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, Genome Biol. 5 (2004) R7.