Contents lists available at ScienceDirect

Genomics Data

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

Data article

RNA-seq analysis for secondary metabolite pathway gene discovery in *Polygonum minus*

Kok-Keong Loke [a], Reyhaneh Rahnamaie-Tajadod [a], Chean-Chean Yeoh [b], Hoe-Han Goh [a,*], Zeti-Azura Mohamed-Hussein [a,b], Normah Mohd Noor [a], Zamri Zainal [a,b], Ismanizan Ismail [a,b]

[a] *Institute of Systems Biology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia*
[b] *School of Biosciences and Biotechnology, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia*

ARTICLE INFO

Article history:
Received 26 October 2015
Accepted 6 November 2015
Available online 7 November 2015

Keywords:
Hybrid assembly
Kesum
Secondary metabolites
Transcriptome

ABSTRACT

Polygonum minus plant is rich in secondary metabolites, especially terpenoids and flavonoids. Present study generates transcriptome resource for *P. minus* to decipher its secondary metabolite biosynthesis pathways. Raw reads and the transcriptome assembly project have been deposited at GenBank under the accessions SRX313492 (root) and SRX669305 (leaf) respectively.

Specifications	
Subject area	Biology, Plant Molecular Biology
Type of data	Transcriptome sequences
Organism/Cell line/tissue	*Polygonum minus* (leaf and root)
Sequencer type	Illumina HiSeq™ 2000 (leaf), Roche 454 GS-FLX (root)
Data format	Raw and processed
Experimental factors	Controlled growth chamber (leaf), experimental plot (root)
Experimental features	RNA-seq dataset for gene discovery in plant
Sample source location	Malaysia
Data accessibility	GenBank accession SRX669305 (http://www.ncbi.nlm.nih.gov/sra/SRX669305) and SRX313492 (http://www.ncbi.nlm.nih.gov/sra/SRX313492)

1. Value of the data

- Current transcriptome datasets greatly improve the previous EST study in *P. minus* [1].
- *P. minus* is a non-model medicinal plant rich in terpenoids bioactive compounds [2].
- Improved transcript repository with increased KEGG pathways coverage provide extensive genetic resource to integrate research between gene expression and metabolite compounds in *P. minus*.
- This data will add to the Polygonum transcriptome resource for understanding secondary metabolite production in this genus.

2. Data

To profile the leaf and root transcriptomes of *P. minus*, RNA-seq short reads were generated from the polyA-enriched cDNA libraries prepared from the total RNAs extracted from the leaf and root tissues. The short reads were filtered, processed, assembled and analyzed as described below. The raw data and assembly project have been deposited at GenBank under the accessions SRX313492 (http://www.ncbi.nlm.nih.gov/sra/SRX313492) and SRX669305 (http://www.ncbi.nlm.nih.gov/sra/SRX669305) for the root and leaf tissues respectively.

3. Experimental design, materials and methods

3.1. Plant materials

Sampling of *P. minus* root and leaf tissues were done from the experimental plot (3° 16′14.63″ N, 101° 41′ 11.32″ E) at Universiti Kebangsaan Malaysia, Bangi. Collected samples were rinsed with distilled water and frozen in liquid nitrogen before stored under −80 °C.

[*] Corresponding author. Tel.: +60389214557; fax: +60389213398.
E-mail address: gohhh@ukm.edu.my (H.-H. Goh).

http://dx.doi.org/10.1016/j.gdata.2015.11.003

Table 1
Statistics of *P. minus* hybrid assembly.

Attributes	Value
Pre-assembly	
Total raw reads	48,615,711
Total processed reads	34,365,872
Post-assembly	
Number of unigenes	108,541
Number of unique transcripts	188,735
N50 (bp)	1009
Size range (bp)	201–12,106

Table 2
Functional annotation of *P. minus* unique transcripts.

Annotation/tools	Number of unique transcripts
Total Transdecoder Peptides	86,295
BLASTX-SwissProt	17,307
BLASTP-SwissProt	29,283
PFAM-TMHMM	13,617
eggNOG	29,004
Gene Ontology (GO)	52,796
SignalP	3715
RNAMMER	9

3.2. Total RNA extraction, quality control, library preparation and RNA-seq

Total RNA was extracted accordingly to protocol reported by Lopez-Gomez [3]. 250 ng of poly(A) RNA was prepared from *P. minus* root sample using PolyATract mRNA isolation kit (Promega, USA) and used as starting material in Roche 454 GS FLX pyrosequencing platform at Malaysia Genome Institute, Malaysia. PCR emulsion was done with long fragment Lib-emPCR amplification for amplicons that are 550 bp or greater in length. The conditions used are as follows: 94 °C for 4 min, 50 cycles of 94 °C for 30 s and 60 °C for 10 min.

Two biological replicates of *P. minus* leaf samples were sequenced using the Illumina HiSeq 2000 sequencing platform. Paired end reads with 90 bp was generated through the standard library preparation protocol implemented by BGI-Shenzhen, P. R. China.

3.3. Transcriptome de novo assembly, annotation and classification

Raw reads were filtered to remove adapter sequences with sequence pre-processing tools, Cutadapt [4] and Trimmomatic [5]. High quality Illumina raw reads with phred score ≥ 25 were kept for assembly. Root 454 reads were clipped to pseudo reads equivalent to that of leaf Illumina short reads of 90 bp with 5 bp overlap using an in-house PHP script (http://gitlab.inbiosis.ws/open-source/rnaseq-utils). These reads were then digital normalized with Khmer protocol (http://khmer.readthedocs.org/en/v1.0/). De novo hybrid assembly of these processed reads was performed with Trinity (release r20140717) [6]. Statistics of the hybrid assembly is showed in Table 1.

Protein coding sequences of unique transcripts were analyzed via Transdecoder which was embedded as a utility script in Trinity pipeline. Standard Trinotate (release r20140708) annotation pipeline (https://trinotate.github.io/) was carried out to annotate the assembled unique transcripts against Swissprot [7], Pfam [8], eggNOG [9], Gene Ontology [10], SignalP [11], and Rnammer [12]. Summary of the annotation is showed in Table 2. Annotated Gene Ontology terms from Trinotate were associated with EC2GO database [13] for KEGG Pathway mapping via KEGG Search & Color Mapper API [14] (Table 3).

Conflict of interest

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

Table 3
Statistics of EC2GO mapped enzymes and KEGG pathway mapping.

Mapping resources	Total mapping entities
GO2EC	482 unique enzymes
KEGG search & color	7037 unique KO, 376 KEGG pathways

Acknowledgments

This research was supported by Fundamental Research Grant Scheme (FRGS/1/2013/SG05/UKM/01/2) from the Malaysian Ministry of High Education (MOHE) and Research University Grant under the Arus Perdana (UKM-AP-BPB-14-2009) from Universiti Kebangsaan Malaysia.

References

[1] N.D. Roslan, J.M. Yusop, S.N. Baharum, R. Othman, Z.-A. Mohamed-Hussein, I. Ismail, N.M. Noor, Z. Zainal, Int. J. Mol. Sci. 13 (2012) 2692–2706.
[2] S.N. Baharum, H. Bunawan, M.a.A. Ghani, W.A.W. Mustapha, N.M. Noor, Molecules 15 (2010) 7006–7015.
[3] R. Lopez-Gomez, M.A. Gomez-Lim, Hortscience 27 (1992) 440–442.
[4] M. Martin, EMBnet.journal 17 (2011) 10–12.
[5] A.M. Bolger, M. Lohse, B. Usadel, Bioinformatics 30 (2014) 2114–2120.
[6] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, M.B. Couger, D. Eccles, B. Li, M. Lieber, Nat. Protoc. 8 (2013) 1494–1512.
[7] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, Nucleic Acids Res. 31 (2003) 365–370.
[8] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, Nucleic Acids Res. (2013) D222–D230.
[9] S. Powell, K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldón, T. Rattei, C. Creevey, M. Kuhn, Nucleic Acids Res. (2013) D231–D239.
[10] C ., Gene Ontology. Nucleic Acids Res. 32 (2004) D258–D261.
[11] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen. Nat. Methods 8 (2011) 785–786.
[12] K. Lagesen, P. Hallin, E.A. Rødland, H.-H. Stærfeldt, T. Rognes, D.W. Ussery. Nucleic Acids Res. 35 (2007) 3100–3108.
[13] E. Camon, D. Barrell, C. Brooksbank, M. Magrane, R. Apweiler, Comp. Funct. Genomics 4 (2003) 71–74.
[14] S. Kawashima, T. Katayama, Y. Sato, M. Kanehisa, Genome Inform. 14 (2003) 673–674.