



ELSEVIER

Contents lists available at ScienceDirect

Data in Brief

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



Data Article

Data of *de novo* assembly and functional annotation of the leaf transcriptome of *Impatiens balsamina*



Lian Chee Foong^a, Anthony Siong Hock Ho^a,
Brandon Pei Hui Yeo^b, Yang Mooi Lim^c, Sheh May Tam^{a,*}

^a School of Biosciences, Faculty of Health and Medical Sciences, Taylor's University, Jalan Taylors, 47500 Subang Jaya, Selangor, Malaysia

^b Formerly of the Faculty of Engineering Computing and Science, Swinburne University of Technology Sarawak Campus, Jalan Simpang Tiga, 93330 Kuching, Malaysia

^c Department of Pre-Clinical Sciences, Faculty of Medicine and Health Sciences, Universiti Tunku Abdul Rahman, Lot PT 21144, Jalan Sungai Long, Bandar Sungai Long, 43000 Kajang, Selangor, Malaysia

ARTICLE INFO

Article history:

Received 19 November 2018

Received in revised form

10 December 2018

Accepted 12 December 2018

ABSTRACT

Impatiens balsamina is both an ornamental and pharmacologically important plant widely distributed in many Asian countries. The leaf of the plant contains many secondary metabolites possessing anti-microbial, anti-tumour and anti-cancer properties. Though there are many phytochemical studies done on the different natural extracts for this plant, not much of genetic information is currently available. This is the first transcriptome of *I. balsamina* leaf using paired-end Illumina HiSeq sequencing which generated 10.79 GB of raw data. Information of pre-processing (reads filtering), *de novo* assembly and functional annotation are presented. This data is accessible via NCBI BioProject (PRJNA505711).

© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license

(<http://creativecommons.org/licenses/by/4.0/>).

* Corresponding author.

E-mail addresses: foonglianchee@sd.taylors.edu.my (L.C. Foong), Anthony.Ho@taylors.edu.my (A.S.H. Ho), yeoph90@gmail.com (B.P.H. Yeo), ymlim@utar.edu.my (Y.M. Lim), shehmay.tam@taylors.edu.my (S.M. Tam).

<https://doi.org/10.1016/j.dib.2018.12.042>

2352-3409/© 2019 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Specifications table

Subject area	Plant Science
More specific subject area	Transcriptomics
Type of data	Table, figure
How data was acquired	Illumina HiSeq. 2000 sequencing platform
Data format	Raw, analysed
Experimental factors	Total RNA was isolated from the leaves of <i>I. balsamina</i>
Experimental features	Total RNA of the pink, multi-petal <i>I. balsamina</i> was extracted from the leaves and used for cDNA library construction, followed by generating paired-end sequencing data using Illumina HiSeq. 2000 system. After pre-processing of the raw reads, Trinity software was used to perform <i>de novo</i> assembly of 9.97 GB clean reads. Functional annotation using BLASTx searches against several online databases including NCBI non-redundant (Nr) protein database, Swiss-Prot, Pfam, InterPro, PROSITE protein databases, as well as Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology (GO).
Data source location	Subang Jaya, Selangor, Malaysia (3.0643° N, 101.6174° E)
Data accessibility	Data is with this article and the raw sequence data has been deposited in the SRA database (PRJNA505711).
Related research article	Lian Chee F, Jian Yi C, Anthony Siong Hock H, Brandon Pei Hui Yeo, Yang Mooi L, Sheh May T. RP-HPLC quantification of naphthoquinones (lawsone and MNQ) and transcriptomics study of genes and pathways involved in naphthoquinones biosynthesis in <i>Impatiens balsamina</i> (Balsaminaceae). Plant Gene, “under review”.

Value of the data

- This is the first *de novo* leaf transcriptome that significantly increased amount of sequence information available for this plant, also useful as reference to other *Impatiens* species.
- The annotated transcripts against KEGG pathways could be useful for researchers working on detailed protein-coding genes of this plant related to transcripts, genes and pathways involved in biosynthesis of secondary metabolites.
- This data will serve as a useful transcriptomic resource for future studies including gene expression, RNAi induction analysis, genomics and functional genomics in *I. balsamina* and other *Impatiens* species.

1. Data

Here, we present the first leaf transcriptomic data of *I. balsamina* generated using Illumina HiSeq. 2000 sequencing technology. The sequencing run generated a total of 10.79 GB (106,867,578 reads) raw data in FASTQ format (has been deposited in the SRA database; PRJNA505711). After pre-processing, *de novo* assemble of the clean reads (9.97 GB, 99,258,630 reads) was performed and the information was summarised in Table 1. The analysis showed that 82.65 % of the total transcripts (75,931 sequences) contained putative coding sequences (CDS). Among the CDS, 58.91 % of CDS had a complete open reading frame which containing defined start and stop codons (Fig. 1). Other than that, 31,200 transcripts were classified as partial CDS. Specifically, 17,289 transcripts were classified as “5 prime_partial len” containing a stop codon and missing start codon, 6564 were grouped as “3 prime_partial len” containing a start codon and lacking stop codon, and 7347 were categorised as “internal len” with missing of both

Table 1
Assembly statistics for the leaf transcriptome of *Impatiens balsamina*.

Summary	Assembled transcripts
Total assembled transcripts	91,873
Total assembled unigenes	66,294
Percent GC	42
N50 contig length	1544
Average assembled transcripts (nt)	943
Total assembled reads (nt)	86,600,995

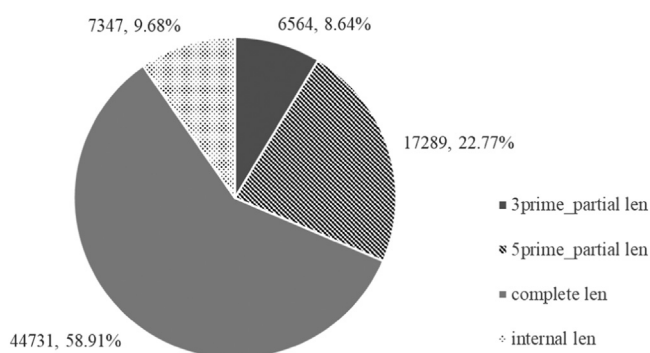


Fig. 1. Types of coding sequences (CDS, with a minimum of 100 amino acids in length) predicted from the assembled leaf transcripts of *Impatiens balsamina*. Complete len = presence of start and stop codons in the open reading frames (ORF) of the CDS; 5 prime_partial len = missing stop codon the ORF; 3 prime_partial len = missing start codon in the ORF; Internal len = neither start nor stop codons were detected in the ORF.

the start and stop codons. A total of 2253 KO IDs was assigned to 24,988 CDS and mapped to 387 KEGG pathways. The output of pathway mapping using KAAS is presented in [Supplementary material S1](#). [Table 2](#) shows an overview of the bioinformatic tools used to analyse the leaf transcriptome of *I. balsamina*.

2. Experimental design, materials, and methods

2.1. Plant material

Cultivated plants of *I. balsamina* (pink, multi-petal form) were obtained from local nursery located in Selangor, Malaysia. The plants were then continuously seed-propagated at the plant growth area, Taylor's University. To minimize sampling variation, leaves were collected and pooled from six plants (age three-months-old). Leaf samples were immediately frozen in liquid nitrogen upon harvesting.

2.2. Total RNA extraction, cDNA library construction and transcriptome sequencing

Total RNA extraction was conducted in triplicates according to an optimized protocol described in [1]. One sample of high-quality intact RNA (RIN > 7.5; A_{260}/A_{280} and A_{260}/A_{230} ratios > 1.9) was then selected for sequencing. Two μg of total RNA from the sample was used in the mRNA-seq library construction. The mRNA was isolated and fragmented to 200 nt for cDNA synthesis. The cDNA was end-repaired, ligated to adapters and PCR-enriched using the NEB Next Ultra RNA Library Prep Kit for Illumina (NEB, USA) according to the manufacturer's protocol. The final library was quantified using a Qubit DNA HS assay and library size determined using a Bioanalyzer High Sensitivity DNA chip.

Table 2
Bioinformatic tools used in the analysis of the leaf transcriptome of *Impatiens balsamina*.

Software/program	Website	Ref.
FastQC v0.11.5	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/	[2]
Trimmomatic v.0.35	http://www.usadellab.org/cms/?page=trimmomatic	[3]
Prinseq v0.20.4	http://prinseq.sourceforge.net/	[4]
Trinity	https://github.com/trinityrnaseq/trinityrnaseq/wiki	[10]
Bowtie v1.1.2	http://bowtie-bio.sourceforge.net/index.shtml	[6]
RSEM v.1.2.11	https://github.com/deweylab/RSEM	[7]
BLAST2GO	https://www.blast2go.com/	[11]
TransDecoder	https://github.com/TransDecoder/TransDecoder/wiki	[8]
KAAS ver. 1.6	http://www.genome.jp/tools/kaas/	[9]
MEGA 7.0.26	https://www.megasoftware.net/	[12]

Sequencing of the final library was performed by Malaysian Genomics Resource Centre Berhad (MGRC) using the Illumina HiSeq. 2000 (Illumina, USA) platform. Paired-end sequencing was performed at 2 × 100 bp per cycle, with 200 cycles.

2.3. Sequence data assembly and bioinformatic analysis

Quality assessment of the reads was performed using FASTQC v0.11.5 [2]. Raw reads were pre-processed using Trimmomatic (version 0.35) [3] and Prinseq (version 0.20.4) [4] to remove the adapter sequences and low quality reads with ambiguous base (N). The sequences with length below 75 were discarded and the remaining sequences were named as 'clean reads'. Trinity software (version 2.2.0) with default parameters was used to *de novo* assemble the clean reads. Transcript and gene IDs were assigned to clean reads according to the default criteria determined by Trinity.

A bioinformatic analysis was performed using the clean reads and the results were reported in [5]. Briefly, bowtie (version 1.1.2) [6] was used to assess the quality of assembled transcripts by mapping the clean reads against assembled transcripts. RSEM (version 1.2.11) software package [7] was used to estimate the abundance of transcripts/unigenes. Functional annotations were accomplished by performing BLASTx searches with an E-value threshold of $\leq 1e^{-5}$ against the NCBI non-redundance protein database, limited to 'green plant (txid 33090)', as well as other established databases such as Swiss-Prot-, Pfam-, InterPro- and PROSITE protein databases. Finally, BLAST2GO program was used to assign gene ontology (GO) terms ($E\text{-value} \leq 1e^{-5}$) to categorise the transcripts.

TransDecoder [8] was used to identify coding sequences (CDS) with open reading frames (ORFs) of at least 100 amino acids in length from the assembled transcripts. Kyoto Encyclopedia of Genes and Genomes (KEGG) Ortholog (KO) assignment and mapping of amino acid sequences to biosynthesis pathways were performed using KEGG automatic annotation server (KAAS; version 1.6) [9], with default threshold bit-score value of 60, single-directional best hit (SBH) method, BLASTx program, and the selected KEGG database included 32 eudicots, two monocots and one basal magnoliophyte. Pathway mapping analysis from KAAS is presented in [Supplementary material S1](#).

Acknowledgements

This work was supported by the Fundamental Research Grant Scheme (FRGS/2/2014/SG05/TAY-LOR/02/1) from the Ministry of Higher Education, Malaysia.

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at <https://doi.org/10.1016/j.dib.2018.12.042>.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at <https://doi.org/10.1016/j.dib.2018.12.042>.

References

- [1] L.C. Foong, S.H. Anthony Ho, Y.M. Lim, S.M. Tam, A modified CTAB-based protocol for total RNA extraction from the medicinal plant *Impatiens balsamina* (Balsaminaceae) for next-generation sequencing studies, *Malays. Appl. Biol.* 46 (2017) 11.
- [2] B. Bioinformatics, FastQC: A Quality Control Tool for High Throughput Sequence Data, Babraham Institute, Cambridge, UK, 2011.
- [3] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics* 30 (2014) 2114–2120.
- [4] R. Schmieder, R. Edwards, Quality control and preprocessing of metagenomic datasets, *Bioinformatics* 27 (2011) 863–864.
- [5] L.C. Foong, J.Y. Chai, S.H. Anthony Ho, P.H. Brandon Yeo, Y.M. Lim, S.M. Tam, RP-HPLC quantification of naphthoquinones (lawsone and MNQ) and transcriptomics study of genes and pathways involved in naphthoquinones biosynthesis in *Impatiens balsamina* (Balsaminaceae), *Plant Gene* (2018) (under review).
- [6] B. Langmead, C. Trapnell, M. Pop, S.L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol.* 10 (2009) 25.
- [7] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, *BMC Bioinforma.* 12 (2011) 323.
- [8] B.J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, et al., De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity, *Nat. Protoc.* 8 (2013) 1494.
- [9] Y. Moriya, M. Itoh, S. Okuda, A.C. Yoshizawa, M. Kanehisa, KAAS: an automatic genome annotation and pathway reconstruction server, *Nucleic Acids Res.* 35 (2007) 182–185.
- [10] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome, *Nat. Biotechnol.* 29 (2011) 644–652.
- [11] S. Götz, J.M. García-Gómez, J. Terol, T.D. Williams, S.H. Nagaraj, M.J. Nueda, et al., High-throughput functional annotation and data mining with the Blast2GO suite, *Nucleic Acids Res.* 36 (2008) 3420–3435.
- [12] K. Tamura, J. Dudley, M. Nei, S. Kumar, Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 4, *Mol. Biol. Evol.* 24 (2007) 1596–1599.