Contents lists available at ScienceDirect

Genomics Data

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

Data in Brief

De novo transcriptome assembly of two contrasting pumpkin cultivars

Aliki Xanthopoulou [a,b], Fotis Psomopoulos [c], Ioannis Ganopoulos [a], Maria Manioudaki [d], Athanasios Tsaftaris [a,b], Irini Nianiou-Obeidat [b], Panagiotis Madesis [a,*]

[a] *Institute of Applied Biosciences, CERTH, Thermi, Thessaloniki 57001, Greece*
[b] *Lab of Genetics and Plant Breeding, School of Agriculture, Forestry and Natural Environment, Aristotle University of Thessaloniki, P.O. Box 261, Thessaloniki GR-54124, Greece*
[c] *Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki 54124, Greece*
[d] *Department of Horticultural Genetics and Biotechnology, Mediterranean Agronomic Institute of Chania Crete, Greece*

ARTICLE INFO

Article history:
Received 30 December 2015
Received in revised form 5 January 2016
Accepted 14 January 2016
Available online 15 January 2016

Keywords:
RNA-Seq
Pumpkin
Contrasting cultivars
Cucurbita pepo

ABSTRACT

Cucurbita pepo (squash, pumpkin, gourd), a worldwide-cultivated vegetable of American origin, is extremely variable in fruit characteristics. However, the information associated with genes and genetic markers for pumpkin is very limited. In order to identify new genes and to develop genetic markers, we performed a transcriptome analysis (RNA-Seq) of two contrasting pumpkin cultivars. Leaves and female flowers of cultivars, 'Big Moose' with large round fruits and 'Munchkin' with small round fruits, were harvested for total RNA extraction. We obtained a total of 6 GB (Big Moose; http://www.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3056882) and 5 GB (Munchkin; http://www.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3056883) sequence data (NCBI SRA database SRX1502732 and SRX1502735, respectively), which correspond to 18,055,786 and 14,824,292 150-base reads. After quality assessment, the clean sequences where 17,995,932 and 14,774,486 respectively. The numbers of total transcripts for 'Big Moose' and 'Munchkin' were 84,727 and 68,051, respectively. TransDecoder identified possible coding regions in assembled transcripts. This study provides transcriptome data for two contrasting pumpkin cultivars, which might be useful for genetic marker development and comparative transcriptome analyses.

Specifications	
Organism/cell line/tissue	Pumpkin (*Cucurbita pepo*)/pool leaves and female flowers
Sex	N.A.
Sequencer or array type	HiSeq2000
Data format	Raw and processed
Experimental factors	N.A.
Experimental features	Leaves and female flowers of two contrasting pumpkin cultivars, 'Big Moose' and 'Munchkin', were harvested for total RNA extraction. Prepared libraries were paired-end sequenced using the HiSeq 2000 system. The obtained data was subjected to de novo transcriptome assembly using Trinity, and coding regions were predicted by TransDecoder.
Consent	N/A
Sample source location	Thessaloniki, Greece, 40°38′37″ N, 22°55′51″ E

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3056882 for pumpkin cultivar 'Big Moose'.

http://www.ncbi.nlm.nih.gov/Traces/sra/?run=SRR3056883 for pumpkin cultivar 'Munchkin'.

2. Experimental design, materials and methods

2.1. Plant materials

Cucurbita pepo (squash, pumpkin, gourd), a world-wide cultivated vegetable of American origin, is extremely variable in fruit characteristics [1] and its genome was recently published in draft format by COMAV's Bioinformatics & Genomics and Cucurbits Breeding groups (https://cucurbigene.upv.es/genome-v3.2/). The small fruit cultivar 'Munchkin' and the large fruit cultivar 'Big Moose' were grown in the experimental stations of Plant Breeding and Genetic Resources institute (ELGO-DEMETER) under standard field conditions. Pest control and water management were performed according to standard practices. Five leaves and female flowers were harvested from a single plant of each cultivar and immediately frozen in liquid nitrogen for further experiments.

2.2. RNA isolation, library preparation and sequencing

Total RNA was extracted from each tissue using the TRIzol® Reagent (Invitrogen, USA) and DNase I was used to remove DNA. We combined equivalent amounts of RNA from each tissue into two pools, one per cultivar. The paired-end library preparation and sequencing were achieved using a NEBNext Ultra RNA library prep kit for Illumina (New England

[*] Corresponding author.

http://dx.doi.org/10.1016/j.gdata.2016.01.006

Table 1
Summary of *de novo* assembled two *Cucurbita pepo* transcriptomes.

Index	Big Moose	Munchkin
Total trinity transcripts	84,727	68,051
Total trinity 'genes'	66,540	56,163
Percent GC	42.44	42.98
Contig N50	1760	1691
Median contig length	590	616
Average contig	1044.21	1037.08
Total assembled bases	88,473,202	70,574,057

Biolabs, Ipswich, MA) and the libraries were sequenced using a HiSeqTM2000 device (Illumina).

2.3. De novo transcriptome assembly, identification protein coding regions, and annotation

We obtained a total of 6 GB and 5 GB sequence data from 'Big Moose' and 'Munchkin' cultivars, respectively, which correspond to 18,055,786 and 14,824,292 150-base reads. In order to consistently apply quality and adapter trimming to the sequences, the cutadapt (https://cutadapt.readthedocs.org/en/stable/) and FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) tools where applied through the Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) wrapper application. After quality assessment, the clean sequences where 17,995,932 and 14,774,486 respectively, which roughly correspond to a loss of ~0.33% in both cases (i.e., high quality sequences). *De novo* transcriptome assembly was performed using Trinity, which uses the de Bruijn graphs algorithm [2]. Detailed information of assembled transcriptome is summarized in Table 1. The numbers of total transcripts for 'Big Moose' and 'Munchkin' were 84,727 and 68,051 and N50 values for 'Big Moose' and 'Munchkin' were 1760 and 1691, respectively. Next, we identified possible protein coding regions within the assembled transcripts using the TransDecoder program implemented in the Trinity software distribution. We identified 74,990 and 64,322 proteins from 'Big Moose' and 'Munchkin', respectively. In summary, the transcriptome sequences of two contrasting cultivars provide a solid foundation for functional genomics studies on pumpkin in the future and will facilitate a better understanding of the molecular mechanisms of fruit morphology.

Conflict of interest

The authors declare that they have no competing interests.

References

[1] A. Xanthopoulou, I. Ganopoulos, A. Tsaballa, I. Nianiou-Obeidat, A. Kalivas, A. Tsaftaris, P. Madesis, Summer squash identification by high-resolution-melting (HRM) analysis using gene-based EST–SSR molecular markers. Plant Mol. Biol. Report. 32 (2) (2014) 395–405.

[2] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29 (7) (2011) 644–652.