Genomics Data 8 (2016) 117–118

Contents lists available at ScienceDirect

Genomics Data

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

De novo transcriptome assembly of *Sorghum bicolor* variety Taejin

Yeonhwa Jo [a,1], Sen Lian [b,1], Jin Kyong Cho [c], Hoseong Choi [a], Sang-Min Kim [d], Sun-Lim Kim [d], Bong Choon Lee [d], Won Kyong Cho [a,c,*]

[a] Department of Agricultural Biotechnology, College of Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Republic of Korea
[b] College of Crop Protection and Agronomy, Qingdao Agricultural University, Qingdao, Shandong 266109, China
[c] The Taejin Genome Institute, Gadam-gil 61, Hoengseong, 25239, Republic of Korea
[d] Crop Foundation Division, National Institute of Crop Science, RDA, Wanju, 55365, Republic of Korea

ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form 2 May 2016
Accepted 3 May 2016
Available online 5 May 2016

Keywords:
RNA-Seq
Sorghum bicolor
Transcriptome
Variety

ABSTRACT

Sorghum (*Sorghum bicolor*), also known as great millet, is one of the most popular cultivated grass species in the world. Sorghum is frequently consumed as food for humans and animals as well as used for ethanol production. In this study, we conducted *de novo* transcriptome assembly for sorghum variety Taejin by next-generation sequencing, obtaining 8.748 GB of raw data. The raw data in this study can be available in NCBI SRA database with accession number of SRX1715644. Using the Trinity program, we identified 222,161 transcripts from sorghum variety Taejin. We further predicted coding regions within the assembled transcripts by the TransDecoder program, resulting in a total of 148,531 proteins. We carried out BLASTP against the Swiss-Prot protein sequence database to annotate the functions of the identified proteins. To our knowledge, this is the first transcriptome data for a sorghum variety derived from Korea, and it can be usefully applied to the generation of genetic markers.

Specifications	
Organism/cell line/tissue	Sorghum (*Sorghum bicolor* variety Taejin)/leaves
Sex	N.A.
Sequencer or array type	HiSeq2000
Data format	Raw and processed
Experimental factors	*De novo* transcriptome assembly of sorghum variety Taejin
Experimental features	Leaves of five sorghum plants (variety Taejin) were harvested for total RNA extraction. A prepared library was paired-end sequenced using the HiSeq 2000 system. The obtained data were subjected to *de novo* transcriptome assembly using Trinity, and coding regions were predicted by TransDecoder. We performed BLASTP against the Swiss-Prot protein database to annotate the identified proteins.
Consent	N/A
Sample source location	Hoengseong, South Korea (37°28′49.6″N 127°58′34.3″E)

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/sra/SRX1715644 for *Sorghum bicolor* variety Taejin.

2. Introduction

Sorghum (*Sorghum bicolor*)—also known as great millet, dura, jowar, or milo—is one of the most popular cultivated grass species in the world. Sorghum is frequently consumed as food for humans and animals as well as used in ethanol production [1]. Sorghum is the fifth most important cereal crop after rice, wheat, maize, and barley. The origin of sorghum is northern Africa; however, sorghum is currently cultivated in tropical and subtropical regions [2]. Because sorghum is native to tropical climates, researchers are developing new cultivars resistant to cold tolerance. In addition, the availability of the genome of sorghum facilitates sorghum research, such as breeding new cultivars and studies on drought-tolerance mechanisms [3]. Sorghum, along with Italian millet and hog millet, has long been cultivated in Korea, and it is currently consumed as a healthy food in Korea. In this study, we performed *de novo* transcriptome assembly for sorghum variety Taejin by next-generation sequencing.

* Correspondig author.
E-mail address: wonkyong@gmail.com (W.K. Cho).
[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.gdata.2016.05.002

Table 1
Summary of *de novo* assembled transcriptomes for *S. bicolor* variety Taejin.

Index	Taejin
Total trinity transcripts	222,161
Total trinity components	95,030
Percent GC	46.54
Contig N50	2374
Median contig length	1100
Average contig	1471.47
Total assembled bases	326,902,359

3. Experimental design, materials, and methods

3.1. Plant materials

Plants for sorghum variety Taejin were grown in a field located in Gadam-ri, Hoengseong-up, South Korea. Leaves from five such plants were harvested and immediately frozen in liquid nitrogen for further experiments.

3.2. RNA isolation, library preparation, and sequencing

Ten leaves collected from five plants were pooled and used for total RNA extraction using the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). For mRNA library preparation, we used a TruSeq RNA Library Prep Kit v2 according to the manufacturer's instructions (Illumina, San Diego, U.S.A.). In brief, the poly-A-containing mRNAs were isolated using poly-T oligo-attached magnetic beads. The first strand of cDNA, followed by a second strand of cDNA, was synthesized from purified mRNAs. End repair was performed followed by adenylation of 3′ ends. Adapters were ligated, and PCR was conducted to selectively enrich DNA fragments with adapters and to amplify the amount of DNA in the library, respectively. The quality control of generated libraries was conducted using a 2100 Bioanalyzer (Agilent, Santa Clara, U.S.A.). The libraries were paired-end sequenced by Macrogen Co. (Seoul, South Korea) using the HiSeq 2000 platform.

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

We obtained 8.748 GB of raw data from sorghum variety Taejin by paired-end sequencing. We conducted *de novo* transcriptome assembly for sorghum variety Taejin using Trinity program ver. 2.0.6, which uses the de Bruijn graphs algorithm [4]. Detailed information on the *de novo* transcriptome assembly is summarized in Table 1. The numbers of total transcripts and components for sorghum variety Taejin were 222,161 and 95,030, respectively. The N50 value was 2374 bp, and the median contig length was 1100 bp. We further predicted coding regions within the assembled transcripts by the TransDecoder program implemented in the Trinity software distribution. As a result, we identified a total of 148,531 proteins. We carried out BLASTP against the Swiss-Prot protein sequence database to annotate the functions of the identified proteins. All but 37,653 proteins were matched to known proteins in Swiss-Prot. Many proteins were homologous to eukaryotes (9643 proteins) followed by bacteria (223 proteins) and viruses (73 proteins). To our knowledge, this is the first transcriptome data for a sorghum variety derived from Korea, and it can be usefully applied to the generation of genetic markers for sorghum species.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgments

This work was carried out with the support of the “Cooperative Research Program for Agriculture Science & Technology Development (Project No. PJ01186102)” conducted by the Rural Development Administration, Republic of Korea. This work is dedicated to the memory of my father, Tae Jin Cho (1946–2015).

References

[1] G. Palmer, Sorghum—food, beverage and brewing potentials. Process Biochem. 27 (1992) 145–153.
[2] J.M. De Wet, J. Harlan, The origin and domestication of sorghum bicolor. Econ. Bot. 25 (1971) 128–135.
[3] A.H. Paterson, J.E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, The sorghum bicolor genome and the diversification of grasses. Nature 457 (2009) 551–556.
[4] 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 (2011) 644–652.