Data in Brief

*De novo* transcriptome assembly of two different apricot cultivars

Yeonhwa Jo, Sen Lian, Jin Kyong Cho, Hoseong Choi, Hyosub Chu, Won Kyong Cho

1. Direct link to deposited data

http://www.ncbi.nlm.nih.gov/sra/SRX1186946 for Apricot cultivar Harcot
http://www.ncbi.nlm.nih.gov/sra/SRX1186893 for Apricot cultivar Ungarische Beste.

2. Introduction

*Apricot (Prunus armeniaca)* belonging to the *Prunus* species is a popular kind of stone fruit tree. Apricot is native to Armenia and is currently cultivated in many countries with climates adaptable for apricot growth. In general, fresh fruits as well as dried apricot are produced. However, the information associated with genes and genetic markers for apricot is very limited. In this study, we carried out *de novo* transcriptome assembly for two selected apricot cultivars referred to as Harcot and Ungarische Beste, which are commercially important apricot cultivars in the world, using next generation sequencing. We obtained a total of 9.31 GB and 8.88 GB raw data from Harcot and Ungarische Beste (NCBI accession numbers: SRX1186946 and SRX1186893), respectively. *De novo* transcriptome assembly using Trinity identified 147,501 and 152,235 transcripts for Harcot and Ungarische Beste, respectively. Next, we identified 113,565 and 126,444 proteins from Harcot and Ungarische Beste using the TransDecoder program. We performed BLASTP against an NCBI non-redundant (nr) dataset to annotate identified proteins. Taken together, we provide transcriptomes of two different apricot cultivars by RNA-Seq.

© 2015 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Shiga, Japan) and the RNeasy Plant Mini Kit (Qiagen, Hilden, Germany). For mRNA library preparation, we used a TruSeq RNA Library Prep Kit v2 according to 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 cDNA followed by second strand cDNA were 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 the 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 protein coding regions, and annotation

We obtained a total of 9.31 GB and 8.88 GB raw data from Harcot and Ungarische Beste, respectively. De novo transcriptome assembly was performed using Trinity, which uses the de Bruijn graphs algorithm [3]. Detailed information of assembled transcriptome is summarized in Table 1. The numbers of total transcripts for Harcot and Ungarische Beste were 147,501 and 152,235, respectively. N50 values for Harcot and Ungarische Beste were 2027 and 2155, respectively. Next, we identified candidate coding regions within the assembled transcripts using the TransDecoder program implemented in the Trinity software distribution. We identified 113,565 and 126,444 proteins from Harcot and Ungarische Beste, respectively. To annotate proteins, we performed BLASTP against the NCBI non-redundant (nr) dataset. Taken together, we provide transcriptomes of two different apricot cultivars by RNA-Seq.

Conflict of interest

The authors declare that they have no competing interests.

Acknowledgment

This work was carried out with the support of “Cooperative Research Program for Agriculture Science & Technology Development (Project No. PJ00976401)” Rural Development Administration, Republic of Korea.

References

[1] F. Geuna, M. Toschi, D. Bassi, The use of AFLP markers for cultivar identification in apricot. Plant Breed. 122 (2003) 526–531.
[2] S. Ercisli, Apricot culture in Turkey. Sci. Res. Essays 4 (2009) 715–719.
[3] 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.

| Table 1 | Summary of de novo assembled two apricot transcriptomes. |
|---------|---------------------------------------------------------|
| Index   | Harcot          | Ungarische Beste |
| Total trinity transcripts | 147,501 | 152,235 |
| Total trinity components  | 71,386  | 69,387  |
| Percent GC                | 41.64   | 41.79   |
| Contig N50                | 2027    | 2155    |
| Median contig length      | 1024    | 1107    |
| Average contig length     | 13,143.1| 1409.00 |
| Total assembled bases     | 193,861.738 | 214,498.725 |