Data in Brief 15 (2017) 454–458

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Long-read transcriptome data for improved gene prediction in *Lentinula edodes*

Sin-Gi Park [a], Seung il Yoo [a], Dong Sung Ryu [a], Hyunsung Lee [a], Yong Ju Ahn [a], Hojin Ryu [b], Junsu Ko [a], Chang Pyo Hong [a,*]

[a] *Theragen Etex Bio Institute, Suwon 16229, Republic of Korea*
[b] *Department of Biology, Chungbuk National University, Cheongju 28644, Republic of Korea*

ARTICLE INFO

Article history:
Received 21 August 2017
Received in revised form
19 September 2017
Accepted 22 September 2017
Available online 27 September 2017

Keywords:
Gene model
Gene prediction
Lentinula edodes
PacBio Single-molecule real-time (SMRT) transcriptome sequencing

ABSTRACT

Lentinula edodes is one of the most popular edible mushrooms in the world and contains useful medicinal components such as lentinan. The whole-genome sequence of *L. edodes* has been determined with the objective of discovering candidate genes associated with agronomic traits, but experimental verification of gene models with correction of gene prediction errors is lacking. To improve the accuracy of gene prediction, we produced 12.6 Gb of long-read transcriptome data of variable lengths using PacBio single-molecule real-time (SMRT) sequencing and generated 36,946 transcript clusters with an average length of 2.2 kb. Evidence-driven gene prediction on the basis of long- and short-read RNA sequencing data was performed; a total of 16,610 protein-coding genes were predicted with error correction. Of the predicted genes, 42.2% were verified to be covered by full-length transcript clusters. The raw reads have been deposited in the NCBI SRA database under accession number PRJNA396788.

Abbreviations: RNA-Seq, whole transcriptome sequencing; GFF, general feature format

* Corresponding author.
E-mail address: changpyo.hong@theragenetex.com (C.P. Hong).

http://dx.doi.org/10.1016/j.dib.2017.09.052
2352-3409/

Specifications Table

Subject area	Biology
More specific subject area	Genomics and Bioinformatics
Type of data	Table, Figure, GFF
How data was acquired	PacBio single-molecule real-time (SMRT) transcriptome sequencing and evidence-driven gene prediction
Data format	Raw, analyzed
Experimental factors	RNA isolation, cDNA library construction and PacBio sequencing
Experimental features	Long-read transcriptome data with variable lengths were generated, and evidence-driven gene prediction was performed based on the data.
Data source location	The monokaryotic B17 strain of *Lentinula edodes* (KCTC46443) was collected from the Korean Collection for Type Cultures (KCTC) in the Republic of Korea (http://kctc.kribb.re.kr/)
Data accessibility	Raw data from this study are available in NCBI's Sequence Read Archive (SRA) database under accession number PRJNA396788 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA396788/)

Value of the data

- The whole-genome sequence of *L. edodes* has been determined with the objective of discovering candidate genes associated with agronomic traits [1], but experimental verification of gene models with correction of gene prediction errors is lacking.
- PacBio long-read transcriptome data integrated with Illumina short-read RNA-Seq data can enhance the accuracy of gene prediction with error correction and support experimental verification.
- Our data will strengthen genome-wide analyses of *L. edodes* by contributing to the identification of targeted genes associated with a trait, transcriptome profiling, and comparative genomics.

1. Data

A total of 5,285,247 long-reads producing 12.61 Gb of sequence data were generated from three RNA libraries of the monokaryotic B17 strain of *L. edodes* that were size-selected for lengths of < 2 kb, 2–3 kb, and 3–6 kb (Table 1). Those reads were clustered into 36,946 transcripts with a cumulative length of approximately 82.1 Mb and an average length of 2.2 kb (Fig. 1). Based on exon-intron boundary information generated by aligning the PacBio long-read (12.6 Gb) and Illumina short-read (3.36 Gb) [1] RNA-Seq data to the draft genome sequence of *L. edodes* [1], a total of 16,610 protein-coding genes were predicted with error correction (Tables 2 and 3). Of those genes, 1344 were newly identified. The transcriptome data supported 92.9% of the predicted gene models (Fig. 2). Moreover, 7005 gene models (42.2%) were verified to be covered by full-length transcript clusters. Homology-based searches indicated that 76.2% of the predicted genes had homology with known genes. Functional annotations were tentatively assigned for 38.3% of these genes. GFF files and annotations of gene models for *L. edodes* are provided in the Supplementary data (Supplementary material 1 and 2).

2. Experimental design, materials and methods

2.1. Materials

The monokaryotic B17 strain of *L. edodes* (KCTC46443) [1] was obtained from the Korean Collection for Type Cultures (KCTC) in the Republic of Korea (http://kctc.kribb.re.kr/).

Table 1
Summary of PacBio long-read transcriptome data in *L. edodes* B17.

	Library size		
	< 2 kb	2–3 kb	3–6 kb
No. of subreads[a]	2,027,562	1,404,810	1,852,875
Total length of subreads (Gb)	3.36	3.49	5.76
No. of reads of inserts	196,775	207,733	351,503
No. of full-length reads	91,513	96,258	150,541
No. of non-full-length reads	82,559	99,023	188,716
No. of filtered short reads	22,703	12,452	12,246
Polished consensus isoforms	12,874	11,223	12,849
Average length of isoforms (bp)	1373	2236	3064

[a] Adapters and artefacts were removed.

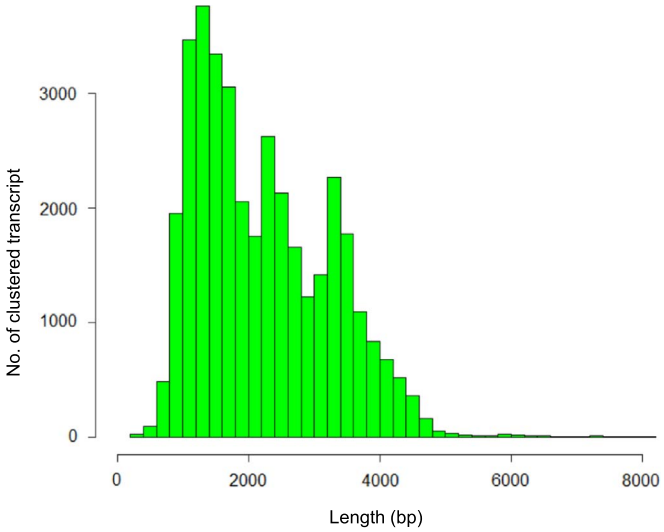

Fig. 1. The length distribution of clustered transcripts.

2.2. RNA extraction and PacBio SMRT transcriptome sequencing

Total RNA from the monokaryotic B17 strain of *L. edodes*, which was cultured in potato dextrose broth liquid medium for 10 days at 25 °C, was extracted using an RNA extraction kit (iNtRon Biotech, Seoul, Korea). cDNA was obtained from the RNA and was size-selected into fractions with the following length ranges: 1–2 kb, 2–3 kb, 3–6 kb, and > 6 kb. SMRTbell template libraries were created from the obtained cDNAs for sequencing on the PacBio RS II system, as recommended by Pacific Biosciences (Palo Alto, U.S.A.). The templates were sequenced via polymerase binding using the DNA/Polymerase Binding Kit P6 v2 primers.

2.3. Long-read transcriptome data clustering, gene prediction and annotation

Long-read transcriptome data clustering was performed using SMRT Analysis software v2.3.0 (https://github.com/PacificBiosciences/SMRT-Analysis) with (i) generation of reads of insert (ROIs), (ii) classification of full-length reads, and (iii) clustering for building consensus sequences.

For gene prediction in the genome of *L. edodes*, AUGUSTUS [2] was used to perform *de novo* prediction with prior gene models trained using GeneMark-ET [3] and exon-intron boundary

Table 2
Summary of gene prediction and annotation updated in *L. edodes* B17.

	This study	Shim et al. (2016)
Protein-coding gene (No.)	18,663	13,426
Unique gene models (No.)	16,610	13,028
Genes with isoforms (No.)	2053	398
Supported by RNA-Seq (No.)	15,263	11,781
Annotated (No.)[a]	12,662	10,700
Average gene length (bp)	1288	1612
Total length of gene models (Mb)	24.05	21.64
Exons		
No. of exons	91,386	77,650
No. of average exons per gene	4.89	5.78
Average exon length (bp)	196	204
Introns		
No. of exons	72,723	64,224
No. of average exons per gene	3.89	4.78
Average exon length (bp)	83	90

[a] Gene models were annotated with homology-based searches.

Table 3
Summary of correction of gene models.

	No. of gene models [1]
Exactly overlapped	7889
Split into ≥ two gene models	4742
Fused with ≥ two gene models	343
Structurally re-predicted	261
Newly found	1344
Predicted in the only previous study	2031

[1] Gene models in the present study were structurally compared with those reported by Shim et al. [1].

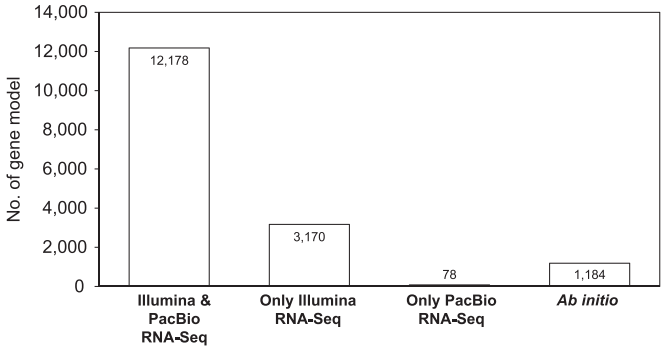

Fig. 2. The distribution of gene models supported by PacBio long-read and Illumina short-read RNA-Seq data.

information predicted by RNA and protein sequence alignments. To generate transcriptome-based evidence, TopHat [4] and GMAP [5] were used for short- and long-read RNA-Seq alignments, respectively. To generate protein-based evidence, homologous protein sequences were collected from the NCBI non-redundant (NR) database, and Exonerate [6] was used for protein sequence alignments to produce protein-based evidence. Predicted genes were searched in the UniProt and NCBI NR

databases using BLASTP [7] with a cut-off *E*-value of 1×10^{-10}. Protein domains were also searched using InterProScan [8] and then assigned to Gene Ontology (GO) terms.

Acknowledgements

This work was supported by the Strategic Initiative for Microbiomes in Agriculture and Food (grant no. 914008-04) from the Ministry of Agriculture, Food and Rural Affairs, Republic of Korea.

Transparency document. Supplementary material

Transparency data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.09.052.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2017.09.052.

References

[1] D. Shim, S.G. Park, K. Kim, W. Bae, G.W. Lee, B.S. Ha, H.S. Ro, M. Kim, R. Ryoo, S.K. Rhee, I.S. Nou, C.D. Koo, C.P. Hong, H. Ryu, Whole genome de novo sequencing and genome annotation of the world popular cultivated edible mushroom, Lentinula edodes, J. Biotechnol. 223 (2016) 24–25.
[2] M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding, Bioinformatics 24 (2008) 637–644.
[3] A. Lomsadze, P.D. Burns, M. Borodovsky, Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm, Nucleic Acids Res. 42 (2014) e119.
[4] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 25 (2009) 1105–1111.
[5] T.D. Wu, C.K. Watanabe, GMAP: a genomic mapping and alignment program for mRNA and EST sequences, Bioinformatics 21 (2005) 1859–1875.
[6] G.S. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison, BMC Bioinform. 6 (2005) 31.
[7] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[8] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, R. Lopez, InterProScan: protein domains identifier, Nucleic Acids Res. 33 (2005) W116–W120.