Data in Brief 17 (2018) 459–462

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Screening and Identification of putative long non coding RNAs from transcriptome data of a high yielding blackgram (*Vigna mungo*), Cv. T9

Pankaj Kumar Singh [a,*], Sayak Ganguli [b], Amita Pal [a]

[a] *Division of Plant Biology, Bose Institute, Kolkata 700054, India*
[b] *Theoretical and Computational Biology Division, AIIST, Palta 743122, India*

ARTICLE INFO

Article history:
Received 26 September 2017
Received in revised form
11 December 2017
Accepted 16 January 2018
Available online 20 February 2018

Keywords:
Blackgram
Long non-coding RNA
Legumes
RNA sequencing data

ABSTRACT

Blackgram (*Vigna mungo*) is one of primary legumes cultivated throughout India, Cv.T9 being one of its common high yielding cultivar. This article reports RNA sequencing data and a pipeline for prediction of novel long non-coding RNAs from the sequenced data. The raw data generated during sequencing are available at Sequence Read Archive (SRA) of NCBI with accession number-SRX1558530

Specifications Table

Subject area	*Biology*
More specific subject area	*Plant molecular biology*
Type of data	*Sequence Data*
How data was acquired	*High throughput sequencing*
Data format	*Raw reads*
Experimental factors	*High Yield Cultivar*

* Corresponding author.
E-mail address: pksingh_12@yahoo.in (P.K. Singh).

https://doi.org/10.1016/j.dib.2018.01.043

Experimental features	*RNA sequencing from total isolated RNA, followed by computational prediction of long non-coding RNA*
Data source location	*Kolkata, West Bengal, India*
Data accessibility	https://www.ncbi.nlm.nih.gov/sra/SRX1558530

Value of the data

- This is the first report of long non-coding RNAs in *Vigna mungo.*
- This study will enable researchers to identify lncRNAs of interest in a high protein yielding legume, *Vigna mungo*.
- This article also contains a pipeline for identification of long non-coding RNAs in *Vigna mungo* an in depth analysis with some adjustments which may pave the way for identification of lncRNAs in other non model plants as well.

1. Data

This works reports the long non-coding RNAs identified in common Indian cultivar of *Vigna mungo* (Blackgram) Cv. T9. This cultivar is widely cultivated in different states of India due to high agronomic yield; however, it is highly susceptible to Mungbean Yellow Mosaic India Virus (MYMIV) infection mediated by the vector whitefly (*Bemisia tabaci).*

2. Experimental design, materials and methods

2.1. RNA isolation and RNA sequencing

Sample preparation for RNA isolation was done as described by Kundu et al. [1]. Total RNA was extracted from prepared sample using Trizol reagent (Invitrogen, Carlsbad, CA) following the manufacturer's instruction, followed by DNase-I treatment (Sigma-Aldrich, USA) and purification using a RNeasy Plant Mini Kit (Qiagen, USA). Qualitative and quantitative assessments of the extracted Total RNA were performed using Agilent 2100 Bioanalyzer (RNA Nano Chip, Agilent). RNA samples were transferred to Genotypic Technologies Pvt. Ltd. (Bangalore, India) for transcript library preparation and for performing high throughput sequencing using Illumina NextSeq. 500 platform. Data generated during this experiment was submitted to Sequence Read Archive (SRA) of National Centre for Biotechnology Information (NCBI) under accession no SRX1558530.

2.2. Bioinformatics analysis and long non-coding RNA prediction

The pipeline shown in Fig. 1 was followed to identify the long non coding RNAs.First raw reads were processed for removal of low quality reads using in house Perl scripts, followed by *de-novo* assembly of transcripts using Trinity [2].*De novo* transcript statistics are provided in Table 1. Processed reads were aligned against assembled transcripts using Bowtie2 [3]. Further BLAST-n [4] was performed against CANTATAdb [5]. Annotated (305 RNAs, Supplementary file 1) and unannotated transcripts (8455 RNAs) were separated. Highest similarities were found with *Glycine max* (65%) (Fig. 2A). Unannotated transcripts were analyzed further, coding potential of transcripts was calculated using CPC Calculator tool [6] and transcripts having low coding potential were selected. Transcripts having length of over 300 bps were selected as suitable candidates for further analyses using TransDecoder. The retained transcripts were again subjected to BLAST nr-Db to establish their non coding character; reads were further searched for similarity against *Vigna mungo* cds (generated via transcriptome sequencing; results unpublished). Remaining 2874 (Supplementary file 2) reads are

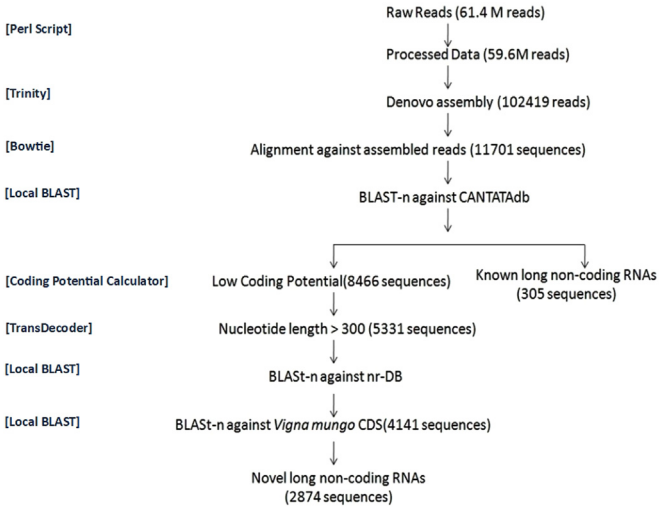

Fig. 1. A flowchart representing novel lncRNA prediction used for this dataset.

Table 1
Assembly statistics for *De novo* assembly.

Transcriptome assembly:	Statistics
Transcripts generated:	102,419
Maximum transcript length:	19,931
Minimum transcript length:	201
Average transcript length	1115.17
Total transcriptss length:	114,214,721 (114.2 MB)
Total number of non-ATGC characters:	0
Percentage of non-ATGC characters:	0
Transcripts > 500 b:	58,588
Transcripts > 1 Kb:	38,336
Transcripts > 10 Kb:	44
n50 value:	1976
n90 value:	436
Number of reads used:	29,522,404
Total number of reads:	29,799,540
Percentage of reads used:	99.07

being proposed as potential novel long non-coding RNAs. This entire pipeline for novel lncRNA prediction is illustrated in Fig. 1.

2.2.1. *Prediction of SSR markers in novel lncRNAs*

Simple sequence repeats were predicted using MISA-MIcroSAtellite identification tool [7]. Ten repeating units for mono nucleotide, 6 repeating units for di nucleotide and 5 repeating units for tri-, tetra-, penta- and hexa nucleotide were chosen as parameters for mining the SSR markers. Details of mined SSRs has been provided in Fig. 2B.

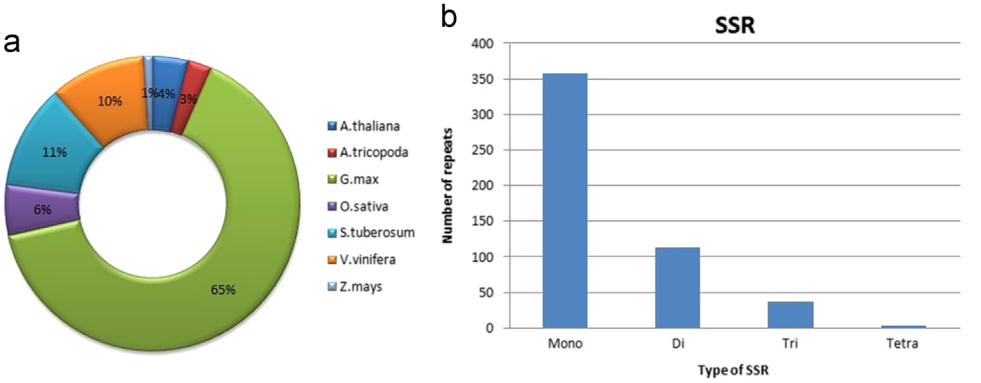

Fig. 2. (a) Doughnut representing similarity of known lncRNAs with lncRNA of different plants from CANTATAdb, (b) represents a histogram of number of SSRs predicted for mono,di,tri,and tetra nucleotide repeats. Majority of SSRs are mono nucleotide repeats.

Acknowledgements

PKS would like to thank Department of Biotechnology, West Bengal 575(sanc)/BT(Estt.)/RD-60/2015 for Senior Research Assistantship and AP is thankful to the UGC for Emeritus Fellowship (F.6-6/2016-17/EMERITUS-2015-17-GEN-7040/(SA-II)). Authors acknowledge Sandor Life Sciences, Hyderabad for their help in data analyses. We are also thankful to the Director, Bose Institute for providing all infrastructural facilities.

Transparency document. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2018.01.043.

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.2018.01.043.

References

[1] A. Kundu, A. Patel, S. Paul, A. Pal, Transcript dynamics at early stages of molecular interactions of MYMIV with resistant and susceptible genotypes oq23f the leguminous host, Vigna mungo, PLoS One 10 (4) (2015) e0124687. http://dx.doi.org/10.1371/journal.pone.0124687.
[2] B.J. Haas, A. Papanicolaou, M. Yassour, et al., De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with trinity, Nat. Protoc. 8 (8) (2013). http://dx.doi.org/10.1038/nprot.2013.084.
[3] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (4) (2012) 357–359. http://dx.doi.org/10.1038/nmeth.1923.
[4] Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman, Basic local alignment search tool, J. Mol. Biol. **215** (1990) 403–410.
[5] M.W. Szcześniak, W. Rosikiewicz, I. Makałowska, CANTATAdb: a collection of plant long non-coding RNAs, Plant Cell Physiol. 57 (1) (2016) e8. http://dx.doi.org/10.1093/pcp/pcv201.
[6] L. Kong, Y. Zhang, Z.-Q. Ye, et al., CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, Nucleic Acids Res. 35 (Web Server issue) (2007) W345–W349. http://dx.doi.org/10.1093/nar/gkm391.
[7] T. Thiel, W. Michalek, R. Varshney, et al., Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.), Theor. Appl. Genet. 106 (2003) 411. http://dx.doi.org/10.1007/s00122-002-1031-0.