Contents lists available at ScienceDirect

Data in Brief

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

Data Article

De novo transcriptome assembly data of the marine bioluminescent dinoflagellate *Pyrocystis lunula*

Damian Menghini, Sylvain Aubry*

Department of Plant and Microbial Biology, University of Zürich, Zollikerstrasse 107, CH-8008, Zürich, Switzerland

ARTICLE INFO

Article history:
Received 20 May 2021
Revised 15 June 2021
Accepted 30 June 2021
Available online 2 July 2021

Keywords:
Dinoflagellates
Pyrocystis lunula
RNAseq
Transcriptome
Bioluminescence
Hybrid assembly

ABSTRACT

Pyrocystis lunula is a unicellular bioluminescing dinoflagellates. While the mechanisms and genes underlying bioluminescence and luciferase synthesis are understood in many bioluminescing clades, it remains unknown in dinoflagellates. We took advantage of merging long and short reads to provide here a *de novo* assembly of *P. lunula* transcriptome. A total of 975 million filtered paired-end reads were obtained and assembled into 155,716 contigs corresponding to putative transcripts that were functionally annotated. This dataset will be valuable for improving our understanding of protist's biology and is accessible via NCBI BioProject (PRJNA727555).

* Corresponding author.
E-mail address: sylvain.aubry@uzh.ch (S. Aubry).
Social media: 🐦 (S. Aubry)

https://doi.org/10.1016/j.dib.2021.107254

Specifications Table

Subject	Algal biology
Specific subject area	Transcriptomics
Type of data	Assembly (fasta file), Table, Figure
How data were acquired	Illumina Novaseq 6000, ONT MinION
Data format	Raw and analyzed
How data were acquired	*Pyrocystis lunula* cultures were harvested at dawn in the dark with or without 2 h shaking treatment.
Description of data collection	Total RNA was extracted from 6 cultures of *P. lunula* using Trizol and subsequent DNAse treatment. cDNA was prepared using the TruSeq sample prep kit and Illumina sequencing performed using 100 bp pair-ended reads. In parallel, mRNA was purified using NebNExt poly(A) mRNA isolation module followed by Oxford Nanopore direct cDNA sequencing kit and sequenced on MinION device.
Data source location	The *P. lunula* strain was obtained from the University of Montreal were grown in culture cabinet at the University of Zürich, Switzerland
Data accessibility	Raw data were deposited in the NCBI SRA database under the Bioproject accession number PRJNA727555, accessible under this link: https://www.ncbi.nlm.nih.gov/bioproject/ PRJNA727555. The assembly transcript contigs and annotation are accessible on Figshare under this link: https://doi.org/10.6084/m9.figshare.14554824.v2

Value of the Data

- We present here the *de novo* assembly of the transcriptome of the bioluminescent dinoflagellate (eukaryotic protists) *Pyrocystis lunula*.
- Dinoflagellates often present very large genomes that remain almost out of reach of current sequencing techniques; therefore, a comprehensive transcriptome is highly valuable to further research.
- The RNAseq has been performed combining long and short reads in order to improve quality of the assembly.
- These data will allow getting more information on the specialized metabolism of dinoflagellates, the genetic basis and regulation of bioluminescence and get more insight into dinoflagellate evolution.

1. Data Description

We present here a *de novo* transcriptome sequencing and assembly of the unicellular dinoflagellate *P. lunula* (Fig. 1A). Many marine protists like *P. lunula* are responsible for the “sea blooming” in various places worldwide [1]. *P. lunula* is a model species used for deciphering circadian rhythms, bioluminescence and photosynthesis, but despite a long history as a model organism, the knowledge about its genomic features remains relatively limited. Sequencing transcriptome of this organism might help shading light on few peculiarities of dinoflagellates more generally, particularly the extent to which transcriptional regulation is actually involved in gene expression in these organisms [2]. A total output of 975 Gb reads was generated from short (Illumina), SRA accession number SRX10783586-SRX10783590 and long (ONT) reads, SRA accession number SRX10783591. In absence of reference genome, reads were filtered and used for the *de novo* transcriptome assembly: 57 % of the total reads were eventually used for assembling the transcript contigs. The resulting transcriptome was 232 Mb size with a GC content of 62 % and a N50 contig length of 1780 bp (Table 1, available at https://doi.org/10.6084/m9.figshare.14554824.v2). We then evaluated the assembled transcriptome by Benchmarking Universal Single-Copy Orthologs (BUSCO, [3]), and shown that 96 % of transcripts were complete BUSCO genes using alveolate genomes as a reference. Transcripts were

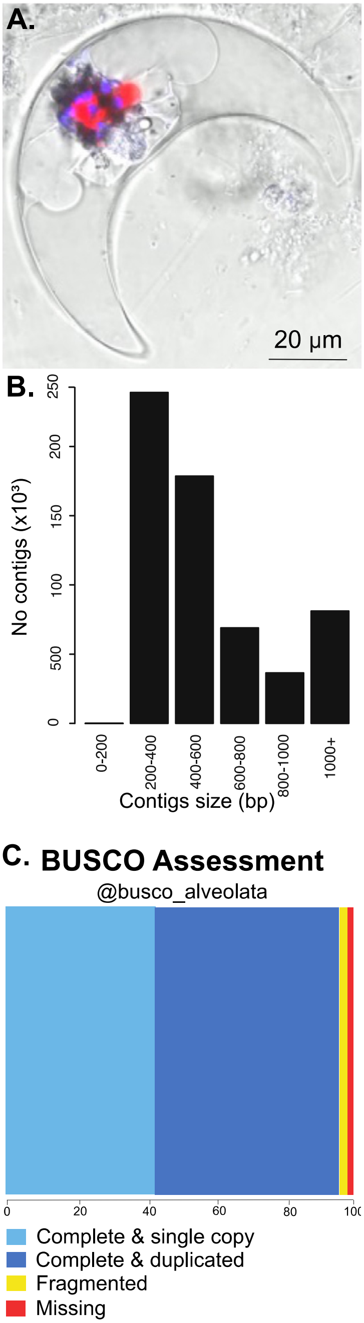

Fig. 1. A. *Pyrocystis lunula* is a unicellular bioluminescing dinoflagellate. Microscope picture showing chlorophyll (red) and luciferase (blue) glowing. B. Contig's size repartition of the assembled transcriptome C. BUSCO assessment of the contigs using alveolate's database from BUSCO.

Table 1
Summary statistics of *de novo* transcriptome assembly for *Pyrocystis lunula* using the combined data of 7 samples.

Transcriptome features	Value
No of contigs	155,716
Largest contig	49,470
Total length	232,137,320
N50	1780
N75	1179
L50	42,944
L75	82,752
GC (%)	61.58

then annotated and classified according to their gene ontology terms using Interproscan and BLAST against *Arabidopsis thaliana* (TAIR 10) proteome (Supplementary Data 1).

2. Experimental Design, Materials and Methods

2.1. Sampling and RNA extraction

Two months old *Pyrocystis lunula* cultures grown in 12 h day/12 h night cycles and kept without mixing at 21 °C and 140 µmol/m^2/s. Total RNA was isolated from shock freezed in liquid nitrogen pelleted cultures using trizol extraction and subsequently treated by DNase digestion step according to the manufacturer's protocol (Qiagen, Germany). The integrity of the RNA was measured on a 4200 TapeStation using the RNA ScreenTape assay (Agilent Technologies, USA).

2.2. Library preparation and sequencing

RNA samples with an RNA integrity number above 8.0 were used for library preparation. A total of 6 cDNA libraries were prepared out of 300 ng total RNA input with the TruSeq RNA Sample Prep Kit v2 (Illumina, USA) according to the manufacturer's protocol. Libraries were pooled and sequenced using an Illumina NovaSeq 6000 sequencing instrument using 100 bp paired-end reads. Sequencing was performed by the Functional Genomic Centre of the University of Zürich.

In parallel, mRNA from one shaked culture was extracted using NEBNext Poly(A) mRNA Magnetic Isolation Module (New England Biolabs) and direct cDNA kit (Oxford Nanopore Technologies). ONT sequencing was performed using a MinION device following the manufacturer's instructions.

2.3. Transcriptome assembly

Illumina paired-end (PE) reads were first quality checked using Fastqc (v0.11.9), MultiQC (v1.9) and FastqScreen (v0.14.1). Afterwards they were adapter trimmed, quality trimmed (4 bp sliding windows from 5' and 3' ends, windows with low quality (<Q20) bases were trimmed) and filtered (average quality of Q20 and above) using fastp (v0.20.0). Processed Illumina reads were mapped to the SILVA rRNA database (release 123) using bowtie2 (v2.4.1) to remove rRNA reads. ONT reads were quality checked using NanoPlot (v1.23.1).

None-rRNA Illumina read pairs and pass filtered ONT reads (quality score of Q7 and above) were assembled using k-mer = 33 into transcripts using rnaspades (v3.14.0) with stranded mode of "–ss rf". Assembled transcriptome was analyzed using BUSCO (v5, [3]), QUAST (v5.0.2, [4]) and EMBOSS (v6.6.0, [5]).

2.4. Functional annotation and gene ontology

Longest ORF per transcript contig was identified using TransDecoder (v5.5.0). Predicted protein sequences were compared against the TAIR10 protein sequences using BLASTP (v2.10.1+). They were also compared to the InterPro database using Interproscan (v5.32-71.0) to obtain gene ontology (GO) and pathway annotation.

CRediT Author Statement

Damian Menghini: Conceptualization, culture and extraction; **Sylvain Aubry:** Conceptualization, Data curation, paper writing, reviewing, editing.

Funding Information

This work was supported by the Swiss National Science Foundation (#31003A_172977).

Ethical Statement

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgements

We thank Prof. David Morse, University of Montreal for providing the *Pyrocystis lunula* culture, and Dr. Maria Domenica Moccia, Dr. Weihong Qi, Dr. Anna Bratus-Neuenschwander from the Functional Genomic Center Zürich for the library preparation, sequencing and bioinformatics analysis.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107254.

References

[1] C. Fajardo, F. Amil-Ruiz, C. Fuentes-Almagro, M. De Donato, G. Martinez-Rodriguez, A. Escobar-Niño, R. Carrasco, J.M. Mancera, F.J. Fernandez-Acero, An “omic” approach to Pyrocystis lunula: New insights related with this bioluminescent dinoflagellate, J. Proteomics. 209 (2019) 103502, doi:10.1016/j.jprot.2019.103502.

[2] S. Roy, M. Beauchemin, S. Dagenais-Bellefeuille, L. Letourneau, M. Cappadocia, D. Morse, The Lingulodinium circadian system lacks rhythmic changes in transcript abundance, BMC Biol. 12 (2014) 107, doi:10.1186/s12915-014-0107-z.

[3] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212, doi:10.1093/bioinformatics/btv351.

[4] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics 29 (2013) 1072–1075, doi:10.1093/bioinformatics/btt086.

[5] P. Rice, I. Longden, A. Bleasby, EMBOSS: The European molecular biology open software suite, Trends Genet 16 (2000) 276–277, doi:10.1016/S0168-9525(00)02024-2.