Data in Brief 35 (2021) 106941

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Dataset of the transcriptomes of *Urechis unicinctus* to identify differentially expressed genes (DEGs) under different temperature and exposure to open air

Xudong Jiao[a,b,*], Jiaxin Shi[c], Song Qin[a,b], Dong Huang[a,d], Yinchu Wang[a,b]

[a] *Key Laboratory of Coastal Biology and Biological Resources Utilization, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai 264003, China*
[b] *Center for Ocean Mega-Science, Chinese Academy of Sciences, Qingdao 266071, China*
[c] *College of Oceanic and Atmospheric Sciences, Ocean University of China, Qingdao 266000, China*
[d] *Agronomy College, Rudong University, Shandong, Yantai 264025, China*

A R T I C L E I N F O

Article history:
Received 16 January 2021
Revised 22 February 2021
Accepted 3 March 2021
Available online 5 March 2021

Keywords:
Urechis unicinctus
Transcriptome assembly
RNA-seq

A B S T R A C T

Urechis unicinctus has a wide range of bioactive polypeptides with high edible, economic and medicinal values. As the key technical breakthrough, the artificial breeding is imperative. However, the seedling transport becomes a primary matter, which indicates the indispensability of realizing how *Urechis unicinctus* responses to various situations. We compared transcriptome of *Urechis unicinctus* under the dry and ultraviolet irradiation treatment and different temperature. The dataset of the organism in response to water-temperature variety was provided by using the Illumina Hiseq X Ten system, which will be helpful to understand the adaptation of *Urechis unicinctus* to changing temperature (low, high and room temperature) and open air (ultraviolet and desiccation). The assembly of the transcriptomes was carried out using the isoform sequencing (Iso-seq) method. The functions of expressed genes were annotated and categorized, while the DEGs were presented.

* Corresponding author.
E-mail address: Xdjiao@yic.ac.cn (X. Jiao).

https://doi.org/10.1016/j.dib.2021.106941

Specifications Table

Subject	Biochemistry, Genetics and Molecular Biology
Specific subject area	Transcriptomics, Genomics
Type of data	Fastq read files
How data were acquired	Illumina HiSeq X Ten Pacific Biosciences (PacBio), Iso-seq method
Data format	Raw sequencing reads (fastq)
Parameters for data collection	Total RNA was collected from 6-month old *Urechis uniconctus* under room temeprature (RT), high temperature (HT), low temeprature (LT), desiccation treatment (DRY) and ultraviolet radiation (UV).
Description of data collection	Total RNA was obtained from 5 groups separately under conditions of UV, DRY and HT, LT and RT, where the RT group was considered as the control one and all groups had 3 parallel experiments.Sequencing was performed according to Illumina HiSeq X Ten. Clean reads were obtained by removing reads containing adapter and low-quality bases and subsequently mapped to the reference spliced by Trinity that is a transcriptome-splicing software combined with 3 separate software modules. The DEGs were analysed by DESeq2.
Data source location	Yantai institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, Shandong, China Harbin Institute of Technology, Weihai, Shangdong, China
Data accessibility	The complete RNA-seq data of *Urechis unicinctus* is available in the NCBI BioProject under accession number (PRJNA603659). Direct URL to data: https://www.ncbi.nlm.nih.gov/bioproject/603659 The sequencing reads of three control groups (RT_1, RT_2, RT_3) used in assembly analysis are available in the NCBI SRA database under accession number: SRX9623339, SRX9623338, SRX9623337 (https://www.ncbi.nlm.nih.gov/sra/SRX9623339 https://www.ncbi.nlm.nih.gov/sra/SRX9623338 https://www.ncbi.nlm.nih.gov/sra/SRX9623337) The sequencing reads of three groups under low temperature(LT_1, LT_2, LT_3) used in assembly analysis are available in the NCBI SRA database under accession number: SRX9623329, SRX9623328 and SRX9623327 (https://www.ncbi.nlm.nih.gov/sra/SRX9623329 https://www.ncbi.nlm.nih.gov/sra/SRX9623328 https://www.ncbi.nlm.nih.gov/sra/SRX9623327) The sequencing reads of three groups under high temperature (HT_1, HT_2, HT_3) used in assembly analysis are available in the NCBI SRA database under accession number: SRX9623340, SRX9623326 and SRX9623325 (https://www.ncbi.nlm.nih.gov/sra/SRX9623340 https://www.ncbi.nlm.nih.gov/sra/SRX9623326 https://www.ncbi.nlm.nih.gov/sra/SRX9623325) The sequencing reads of three ultraviolet groups (UV_1, UV_2, UV_3) used in assembly analysis are available in the NCBI SRA database under accession number: SRX9623336, SRX9623335 and SRX9623334 (https://www.ncbi.nlm.nih.gov/sra/SRX9623333 https://www.ncbi.nlm.nih.gov/sra/SRX9623324 https://www.ncbi.nlm.nih.gov/sra/SRX9623323) The sequencing reads of three desiccation groups (DRY_1, DRY_2, DRY_3) used in assembly analysis are available in the NCBI SRA database under accession number: SRX9623332, SRX9623331 and SRX9623330 (https://www.ncbi.nlm.nih.gov/sra/SRX9623332 https://www.ncbi.nlm.nih.gov/sra/SRX9623331 https://www.ncbi.nlm.nih.gov/sra/SRX9623330)

Value of the Data

- These data show RNA-seq results of *Urechis unicinctus* under ultraviolet and desiccation treatments and changing temperature, providing new insights into the biological pathways of autolytic phenomena.
- These data are useful resources for scientific communities working on transcriptome of *Urechis unicinctus* even invertebrates but also on animal stress biology to understand specific and common stress response pathways.
- Functional analysis data can be used in future studies to anticipate the biological pathways of *Urechis unicinctus* when the temperature changes or being exposure to open air.

1. Data Description

Total RNA was extracted from five groups separately under conditions of ultraviolet, desiccation and high, low and room temperatures. SMRT-bell libraries were constructed after the amplification of optimized polymerase chain reaction (PCR) and sequenced via the PacBio, Iso-seq Sequel and the Illumina Hiseq X Ten platform. However, the single-base error rate was irregular so that multiple corrections were necessary. LoRDEC [1], a software with high precision, corrected the data of three generation sequecing from PacBio with the technique of hybrid error correction. The comparison consequence are displayed in Table 1. We used the Illumina HiSeqTM to produce the raw image data files. These files were transferred into sequenced reads by CASAVA base calling analysis and stored into FASTQ format [2]. The clean data (829,237,442 bp) was obtained by filtering against the NGS QC Toolkit v2.3.3. The primary procedures were removing the adapters and eliminating low-quality bases. 311,403 unigenes without redundancy, with an average length of 1625.39 bp, were annotated and classified by function. We utilized Trinity to splice these clean reads in order to form the reference transcriptome (ref), after which all clean reads were mapped to the ref through RSEM [3]. The result was approximately 101 million RNA-seq reads (Table 2). After finishing the redundancy removing, this assembly was annotated by NCBI-Nr protein database according to different functions. We used readcounts to proceed the analyze of DEGs. DESeq2 were adopted since simples had biological duplication, where the standard of filting was padj$<$0.05 and |log2FoldChange|$>$1. Additionally, binomial distribution method was used to perform independent statistical hypothesis testing, which tended to lead to high overall false positives. Thereby, we needed to correct the p-value obtained from the original hypothesis test.

Table 1
Statistics of length distribution before and after transcript corrections.

Sample	Type	Total nucleotides	Total_number	Mean length	Min length	Max length	N50	N90
UU2019	Before correct	505,187,047	216,918	2329	177	14,121	2518	1493
UU2019	After correct	504,952,370	216,918	2328	176	14,194	2516	1492

Sample: the name of the sample.
Type: the state of correction.
Total_nucleotides: the number of bases of the consensus.
Total_number: the number of the consensus.
Mean length: the average length of the consensus.
Min length: the minimum length of the consensus.
Max_length: the maximum of the consensus.
N50/N90: the total length of the consensus after being ranked in order of length and added up the length until it is no less than 50% or 90% of the consensus.

Table 2
Read alignment summary of transcriptomes of *Urechis unicinctus* under desiccation, ultraviolet and high, low and room temperature.

Sample	Raw Reads	Clean reads	Clean bases	Error(%)	Q20(%)	Q30(%)	GC(%)	Total mapped
DRY_1	49,895,240	48,303,224	7.25 G	0.02	98.36	95.09	47.73	43,022,064(89.07%)
DRY_2	59,168,576	57,459,078	8.62 G	0.02	98.33	95.00	47.13	50,889,032(88.57%)
DRY_3	53,897,626	51,886,052	7.78 G	0.02	98.30	94.98	47.04	45,759,512(88.19%)
UV_1	56,072,498	54,080,562	8.11 G	0.02	98.20	94.73	46.58	47,364,024(87.58%)
UV_2	59,620,598	58,202,014	8.73 G	0.02	98.36	95.09	47.26	51,773,516(88.95%)
UV_3	62,171,266	60,422,202	9.06 G	0.02	98.39	95.15	47.18	53,806,486(89.05%)
RT_1	46,423,864	45,102,198	6.77 G	0.02	98.16	94.64	47.41	39,850,676(88.36%)
RT_2	61,925,716	60,173,960	9.03 G	0.02	98.13	94.54	47.08	53,398,352(88.74%)
RT_3	52,836,082	51,009,134	7.65 G	0.03	97.91	94.07	46.86	45,008,268(88.24%)
HT_1	60,880,848	59,486,478	8.92 G	0.02	98.37	95.11	46.92	52,707,958(88.60%)
HT_2	63,003,754	61,551,050	9.23 G	0.02	98.35	95.10	47.18	54,330,180(88.27%)
HT_3	57,418,292	55,689,904	8.35 G	0.02	98.33	95.01	47.05	49,213,370(88.37%)
LT_1	51,508,474	50,009,866	7.5 G	0.03	97.47	93.07	46.87	44,093,334(88.17%)
LT_2	58,596,068	57,487,684	8.62 G	0.03	97.43	92.96	46.92	50,676,846(88.15%)
LT_3	59,399,558	58,374,036	8.76 G	0.02	98.43	95.30	47.39	51,741,846(88.64%)

Q20, Q30: Proportion of bases with Qphred >20, 30 (Qphred=−10log10(e)).
Raw reads: Original data from sequencing.
Clean Bases: Clean read numbers multiply read length (saved in G unit).
Clean Bases: Clean read numbers multiply read length (saved in G unit).
Error: Average sequencing error rate, calculated through Qphred= −10log10(e).
GC: Propotion of G and C in total bases.

2. Experimental Design, Materials and Methods

2.1. Animal materials and experimental design

We collected *Urechis unicinctus* in LaiShan Bay, BoHai, China (37°27′N, 31°30′E). Fifty individuals of 6-month old *Urechis uniconctus* (average weight: 2.12 g; average length: 3.05 cm) were acclimated at 20 °C in flowing fresh seawater for 2 weeks. They were randomly divided into five groups: Group 1, named RT, was cultured in seawater at a temperature of 20 °C for 2 h; Group 2, named DRY, were cultured without water for 2 h; Group 3, named UV, were cultured in ultraviolet irradiation for 2 h. Group 4 was named LT, where the temperature was instantly decreased to −20 °C and lasted for 120 min; and the temperature of Group 3 (HT) was raised to 28 °C rapidly and lasted for 120 min. Three individuals of each group were randomly sampled and instantly frozen in the liquid nitrogen.

2.2. Total rna extraction, library preparation and sequence

We use the method of TRIzol (Invitrogen, Carlsbad, USA) [4] to extract the total RNA from each mixed sample and the Nanodrop ($OD_{260/280}$ ratio) was used to detect the purity of RNA [5,6]. The Qubit and Aglient 2100 were used to check its concentration and integrity according to the manufacture's protocol. A total of 3 μg mRNA per sample, enriched by Oligo(dT) magnetic beads, was reverse-transcribed through SMARTer® PCR cDNA synthesis kit (Clontech, Mountain View, USA).

Large fragments (>4 kb) double-strand cDNA were selectively used to construct the SMRTbell library and sequenced on the PacBio Sequel platform after repairing DNA damage, end blunting and adapter ligation. In addition, the ployadenylated RNA was broken into short fragments (~200 bp). The double-strand cDNA, which was synthesized with random hexamers after the first-strand preparation and was purified and repaired at the end. The libraries with effective concentration (> 2 nM) were sequenced on the Illumina HiSeq X Ten platform.

Ethics Statement

Each of the procedures that were used to handle and treat the *Urechis unicinctus* during this study was in the accordance with the Animal Management Regulations of China, revised on March 1, 2017, No. 676.

CRediT Author Statement

Xudong Jiao: Conceptualization, Project administration; **Jiaxin Shi:** Data curation, Writing - Original draft preparation; **Song Qin:** Validation, Investigation and Supervision; **Dong Huang:** Writing - Reviewing; **Yinchu Wang:** Data submission and Editing.

Declaration of Competing Interest

The authors declare that they have no competing financial interests, which could influence the work reported in this article.

Acknowledgments

This project was supported by National Key Technology R&D Program of China (No.2018YFA0903000); Youth Innovation Promotion Association, CAS; Regional Demonstration of Marine Economy Innovative Development Project (Yantai, 2020, YHCX-SW-L-202004); Yantai Science and Technology development Program (No.2021YT06000426).

References

[1] L. Salmela, E Rivals, LoRDEC: accurate and efficient long read error correction, Bioinformatics 30 (24) (2014) 3506–3514.
[2] S. Andrews, FastQC A Quality Control Tool for High Throughput Sequence Data, 2010.
[3] B. Li, C Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics (2011) (RSEM), doi:10.1186/1471-2105-12-323.
[4] J.A. Ridgeway, A.E. Timm, Comparison of RNA isolation methods from insect larvae, J. Insect Sci. 14 (2014), doi:10.1093/jisesa/ieu130.
[5] A. Rodríguez, H. Duyvejonck, J.D. Van Belleghem, T. Gryp, V.S. Leen, S. Vermeulen, et al., Comparison of procedures for RNA-extraction from peripheral blood mononuclear cells, PLoS ONE 15 (2) (2020) e0229423, doi:10.1371/journal.pone.0229423.
[6] A. Rodríguez, M. Vaneechoutte, Comparison of the efficiency of different cell lysis methods and different commercial methods for RNA extraction from 0RW1S34RfeSDcfkexd09rT2Candida albicans1RW1S34RfeSDcfkexd09rT2 stored in RNAlater, BMC Microbiol. 19 (2019), doi:10.1186/s12866-019-1473-z.