Data in Brief 36 (2021) 107032

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Draft genome sequence data of *Urechis unicinctus*, a marine echiuroid worm

Xudong Jiao[a,b], Jiaxin Shi[c], Song Qin[a,b], Donghui Zhao[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] *School of Marine Science and Technology, Harbin Institute of Technology, Weihai 264209, China*

ARTICLE INFO

Article history:
Received 3 January 2021
Revised 23 March 2021
Accepted 30 March 2021
Available online 2 April 2021

Keywords:
Urechis unicinctus
Marine worm
Genome annotation
Orthologous groups

ABSTRACT

Urechis unicinctus is mainly distributed in Japan, north Korea and the Yellow Sea and the coast of Bohai Bay in China, and its nutrition is rich. The body of *Urechis unicinctus* contains many types of bioactive polypeptides, such as plasmin and tachykinin, which hold high economic and medicinal values. Therefore, the study of *Urechis unicinctus* has great significance. But the genome of *Urechis unicinctus* remains unavailable till now. To further understand the evolution of *Urechis unicinctus* and determine more effective application of it, we assembled the first draft genome sequence and the assembly of *Urechis unicinctus*. The dataset can be assessed from the BioProject at NCBI (https://www.ncbi.nlm.nih.gov/bioproject/?term=Urechis *unicinctus*).

* Corresponding author at: Key Laboratory of Coastal Biology and Biological Resources Utilization, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai 264003, China.
E-mail address: ycwang@yic.ac.cn (Y. Wang).

https://doi.org/10.1016/j.dib.2021.107032

Specifications Table

Subject	Biochemistry, Genetics and Molecular Biology
Specific subject area	Genomics
Type of data	Tables
How data were acquired	High-throughput sequencing (Illumina HiSeq 4000)
Data format	Raw sequencing reads (fastq)
Parameters for data collection	Total DNA was collected from 6-month old Urechis uniconctus by Laishan Bay, Yantai, Shandong, China
Description of data collection	Sequencing was performed according to Illumina HiSeq 4000. We used K-mer to analyse the size of the genome and the Soapdenovo software to implement assembly.
Data source location	Yantai institute of Coastal Zone Research, Chinese Academy of Sciences, Yantai, Shandong, China
Data accessibility	The complete genome sequence of *Urechis unicinctus* is available in the NCBI BioProject under accession number (PRJNA603659). URL: https://www.ncbi.nlm.nih.gov/bioproject/603659 The sequencing reads used in assembly analysis are available in the NCBI SRA database under accession number: SRX8117413 (https://www.ncbi.nlm.nih.gov/sra/?term=SRX8117413).

Value of the Data

- The data reports the first draft genome of Urechis unicinctus. This is the first genome project of a species of the Urchidae, which can be used as a reference for the analysis of other species.
- The dataset is a useful resource for researchers occupied on genome survey of other species of Urchidae, even invertebrates, and also on gene identification, microsatellite markers, and other genome elements.
- The draft genome is helpful to repeat the genome regions and can be used for the comparative analysis in evolutionary studies. Thereby, some biological properties of Urechis unicinctus can be expected to be discovered.

1. Data Description

The draft genome (1130,128,429 bp) recovered 87.32% of the estimated genome size (1319,380,000 bp) distributed in 3742,050 contigs, and the amount of which is 75.3 G. The sequence data can be accessed at: https://www.ncbi.nlm.nih.gov/sra/?term=SRX8117413.

Detailed features of the draft genome sequence of four libraries are shown in Table 1. The total length of reads assembled de novo was 1152,060,386 bp, and there were 3551,270 scaffolds

Table 1
Statistical data summary of the draft genome.

Library name	Raw paired reads	Raw Base(bp)	Effective Rate(%)	Error Rate(%)	Q20(%)	Q30(%)	GC Content(%)
DES03138-V_L8	55,297,767	16,589,330,100	99.92	0.03	95.84	90.41	40.27
DES03138-V_L1	65,174,254	19,552,276,200	99.78	0.03	96.61	92.12	40.30
DES03139-V_L1	68,791,280	20,637,384,000	99.91	0.03	96.69	92.29	40.27
DES03139-V_L8	61,643,011	18,492,903,300	99.92	0.03	95.97	90.69	40.25

Q20, Q30: Proportion of bases with Qphred >20, 30 (Qphred=−10log10(e)).
Raw reads: Original data from sequencing.
Raw Base: Raw read numbers multiply read length (saved in bp unit).
Effective Rate: The ratio of clean data after filtering to raw data.
Error Rate: Average sequencing error rate, calculated through Qphred = −10log10(e).
GC: Propotion of G and C in total bases.

Table 2
The length (including total, maximum, N50 and N90) and total number of sequenced reads by contigs and scaffolds.

Title	Total_length (bp)	Total_number	Max_length (bp)	N50_length (bp)	N90_length (bp)
Contig	1130,128,429	3742,050	36,805	469	121
Scaffold	1152,060,386	3551,270	38,471	529	124

of at least 100 bp with the largest scaffold 38,471 bp, an N50 length of 529 bp and an N90 length of 124 bp (Table 2). The average covering of the total assembly was 57.137-fold. The *Urechis unicinctus* genome sequence would be a useful resource for the future molecular evolutionary analysis.

2. Experimental Design, Materials and Methods

2.1. Genomic DNA sampling and sequencing

In this study, *Urechis unicinctus* were collected from Laizhou Bay of the Bohai Sea in China. The total DNA was extracted from *Urechis unicinctus* using the method of phenol-chloroform extraction [1]. The sample DNA was broken to fragment by Covaris Ultrasonic Breaker, after which the end-repair, poly-A, adaptor addition, selecting fragment and PCR were proceeded successively. In particular, the concentration of the extracted DNA was determined by Qubit Fluorometer and its integrity and purity was detected by agarose gel electrophoresis [2,3]. A 250–300 bp library was constructed using Chromium^{TM} Genome Reagent Kits, and the sequencing was performed using an Illumina HiSeq 4000.

2.2. Sequence quality control

Removing the adapters and low-quality bases are crucial to obtain clean reads with high quality. The base mass value of the original data of second-generation transcriptome sequencing is automatically given by the sequencer, all of which are ASCII codes. The corresponding ASCII value of each character minus 33 is the sequencing mass value of the base. Quality control was performed using NGQC which is an independent research and development software by Novogene.

Taxonomic assignation was performed on 10,000 randomly selected reads using Blast software (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and NT library in order to detect potential contamination.

2.3. Sequencing assembly

Soapdenovo version 2.04 software was used for the assembly. The result was then assembled into scaffold using Kmer = 41 which the company selected based on project experience in order to evaluate the size of the genome. The assembly results are shown in Table 3.

Table 3
The estimation results of genome size and heterozygosity rate using version jellyfish 2.2.7 for K-mer analysis.

Kmer	Depth	n kmer	Genome size (M)	Revised Genome size (M)	Heterozygous rate (%)	Repeat rate (%)
17	46	62,034,867,589	1348.58	1319.38	1.26	52.64

Ethics Statement

Each of the procedures that were used to handle and treat the *Urechis unicinctus* during this study was approved by the Key Laboratory of Coastal Biology and Biological Resources Utilization, Yantai Institute of Coastal Zone Research, Chinese Academy of Sciences prior to the initiation of the study and the Animal Management Regulations, 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; **Donghui Zhao:** Data submission; **Yinchu Wang:** Writing - Reviewing 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); Yantai Science and Technology Development Program (No. 2021YT06000426).

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107032.

References

[1] S. Chandra, A.K. Varshney, S. Sinha, S. Nehra, N. Mathur, A rapid and efficient DNA Extraction method for PCR based assays in activated sludge, J. Plant Sci. Res. 26 (2) (2010) 201–206.
[2] J.T. Corthell, Agarose Gel Electrophoresis, Basic Molecular Protocols in Neuroscience: Tips, Tricks, and Pitfalls, Academic Press, 2014, pp. 21–25.
[3] High-throughput analysis of lignin by agarose gel electrophoresis, J. Agric. Food Chem. (2020).