Data in Brief 50 (2023) 109449

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Data on SSR markers and SNPs filtered from transcriptome of *Parvocalanus crassirostris*

Nazima Habibi[a,*], Saif Uddin[a], Montaha Behebehani[a], Mohd Wasif Khan[b], Nasreem Abdul Razzack[a], Faiz Shirshikhar[a]

[a] Environment and Life Sciences Research Centre, Kuwait Institute for Scientific Research, Kuwait
[b] Department of Biochemistry and Medical Genetics, University of Manitoba, Winnipeg, MB, Canada

ARTICLE INFO

Article history:
Received 6 February 2023
Revised 20 June 2023
Accepted 21 July 2023
Available online 28 July 2023

Dataset link: RNA Seq of Parvocalanus crassirostris (Original data)
Dataset link: Transcriptome Assembly of Parvocalanus crassirostris (Original data)

Keywords:
Transcriptome
Single nucleotide polymorphism
Simple sequence repeat
Indels
Primers

ABSTRACT

Calanoid copepod populations are being severely affected due to the effects of ocean acidification (OA) and ocean warming (OW). These marine organisms are the most abundant primary consumers contributing significantly in the marine food web. Any effect on the abundance and diversity of copepods due to climate change is likely to have serious implications on the marine ecosystem functioning. Molecular studies that play a vital role in assessing the genetic changes under the influence of environmental imbalances are completely lacking for this species. Here we report the genetic variations in three generations of copepods through transcriptome sequencing. RNA sequencing was performed on an Illumina HiSeq platform employing the 2 × 100 bp paired-end chemistry. Approximately, 10GB of data was obtained for all the samples. The raw sequences were assembled through Trinity 2.6.6 and mined for single nucleotide polymorphisms (SNPs) and simple sequence repeats (SSRs). MIcroSAtellite Identification Tool (MISA) was used for SSR detection and Primer 3 (v 3.0) was utilized to design short oligonucleotide primers (18-20 mers). A total of 15,222 SSRs were identified and 28,944 primer pairs were designed against these motifs. The transcriptome possessed 413,890 SNPs at a frequency of 2.8 per kb. The newly discovered SSRs and SNPs could act as

* Corresponding author.
E-mail address: nhabibi@kisr.edu.kw (N. Habibi).
Social media: @nazima_Habibi (N. Habibi), @saif_ud_din (S. Uddin)

https://doi.org/10.1016/j.dib.2023.109449

genetic markers for future studies on genetic diversity and conservation for *Parvocalanus crassirostris*.

Specifications Table

Subject	Marine Biology
Specific subject area	Genomics
Type of data	Tables, figures, raw sequencing reads, microsatellite motif file, SSR primer file, SNPs and indel file
How the data were acquired	Paired-end (2 × 100 cycles) sequencing on Illumina HiSeq 2500
Data format	Raw, analysed, filtered
Description of data collection	*Parvocalanus crassirostris* isolated from Kuwait's territorial waters were cultured under laboratory conditions. Three generations of copepods were retrieved at 1 month interval and used for RNA isolation. High quality RNA was subjected to whole transcriptome sequencing. Raw sequences were *de novo* assembled and mined for simple sequence repeats, single nucleotide polymorphisms and indels.
Data source location	• Institution: Kuwait Institute for Scientific Research • City/Town/Region: Shuwaikh, Kuwait • Country: Kuwait • Latitude and longitude (29°18′50.67″N; 47°29′30.30″E)
Data accessibility	Repository name: National Centre for Biotechnology Information Data identification number: PRJNA902692 (SRX18292872, SRX18292870, SRX18292869). Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/?term=parvocalanus Transcriptome Assembly: GKHX00000000.1; Repository name: Figshare Direct URL to data: https://figshare.com/s/1916c6b93df6ce1d6016

Value of the Data

- Newly developed SSR markers of *Parvocalanus crassirostris* will be useful for molecular and population genetics studies of this calanoid copepod species.
- These molecular markers can be used for the assessment of genetic diversity of the species and other closely related copepods.
- The data will be beneficial to conservation biologists for management of copepod populations.
- SNPs identified the current dataset provide an extensive set of genetic markers for copepod for future studies on phylogeny and genetic mapping

1. Objective

The data presented in this article was generated with an aim to obtain the first genomic resource of the calanoid copepod *Parvocalanus crassirostris* that are severely impacted due to the unwanted effects of ocean acidification and ocean warming.

2. Data Description

We present the data of RNA sequencing of *Parvocalanus crassirostris*, a calanoid copepod. Raw data of 10Gb was generated and assembled into a fasta file of 154.7 Mb. The sequences were deposited on National Centre for Biotechnology Information (NCBI) under the accession number

PRJNA902692 and the assembly with an accession number GKHX00000000.1. The transcriptome of *P. crassirostris* contained 15,222 SSR regions distributed within 12,644 sequences. A total of 1,973 sequences possessed single SSRs, whereas 1,811 sequences had SSRs in compound formation. The total number of sequences examined for SSRs was 249,255. The size of the sequence was 147,561,992 bp. GC content of assembled sequences was *ca.* 50% (Table 1).

Table 1
Summary of SSR regions examined in the assembled transcriptome of Parvocalanus crassirostris.

Total number of sequences examined	249,255
Total size of examined sequences (bp)	147,561,992
Total number of identified SSRs	15,222
Number of SSR containing sequences	12,664
Number of sequences containing more than 1 SSR	1,973
Number of SSRs present in compound formation	1,811

Further investigations on the motif types revealed, dinucleotides were the largest in number (5,393; 35.42%) followed by trinucleotides (2,454; 16%), tetranucleotides (1,348; 9.0%), pentanucleotides (306; 2.0%), hexanucleotides (100; <1.0%). The frequency distribution of all the tetra- to hexa- nucleotides was below 1.0% (Fig. 1). The top ten repeat motifs (all dinucleotides) in descending order were TG (1,105; 7.26%) > AC (972; 6.39%) > GT (774; 5.08%) >CA (627; 4.12%) > AG (434; 2.85%) > TC (379; 2.49%) >GA (338; 2.22%) > CT (277; 1.82%) > AT (224; 1.47%) and TA (156; 1.02%) (Fig. 2a). The SSR loci pair of AC/GT-22.8% was maximum. Other pairs were AG/CT-9.38%, AAC/GTT-3.00%, AGG/CTT-2.55%, AT/AT-2.50%, ATC/ATG- 2.40%, AGC/CTG- 2.33%, AAAC/GTTT-2.23%, ACC/GGT-1.45 % and CCG/GCC-1.41% (Fig. 2b). The SSR loci length ranged from 5 to 74, only the top ten are shown in Fig. 2c. Maximum repeat numbers were 5 (14.1%). The number of SSR containing sequences decreased as the repeat number of motifs increased. Although 78% of SSRs were located in undetermined region of the transcriptome, however 2.22%, 3.54% and 4.01% were present in the cds, utr3 and utr5 regions, respectively. The SSRs in the sequences existed in a complex repetitive, mono base repetitive, di base repetitive or tri base repetitive pattern. All the 13,412 SSR motifs were subjected to primer designing employing the standard parameters. Primer pairs were obtained for 9,648 SSR motifs. Three primer pairs were designed for each motif, hence the total number of primer pairs summed up to 28,944. All the

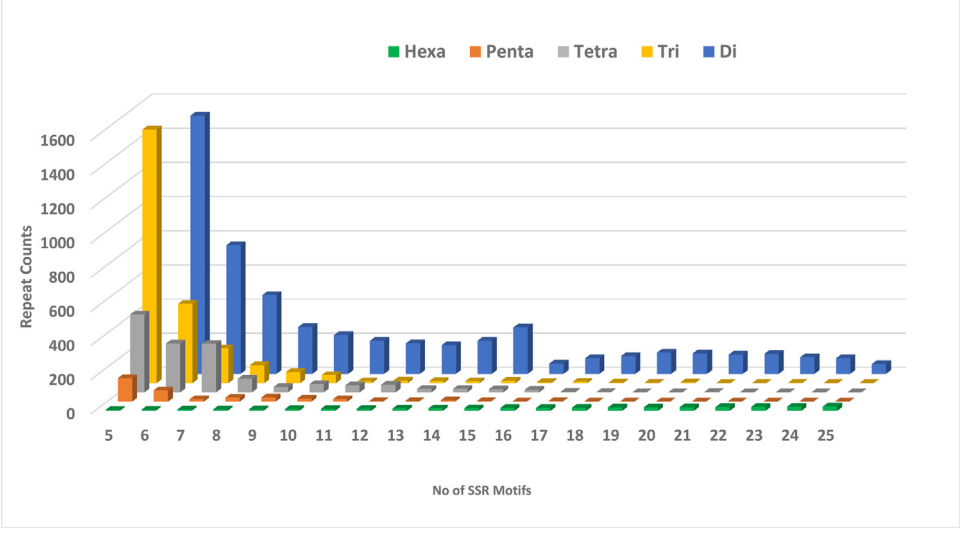

Fig. 1. SSR motifs mined from transcriptome of *Parvocalanus crassirostris.*

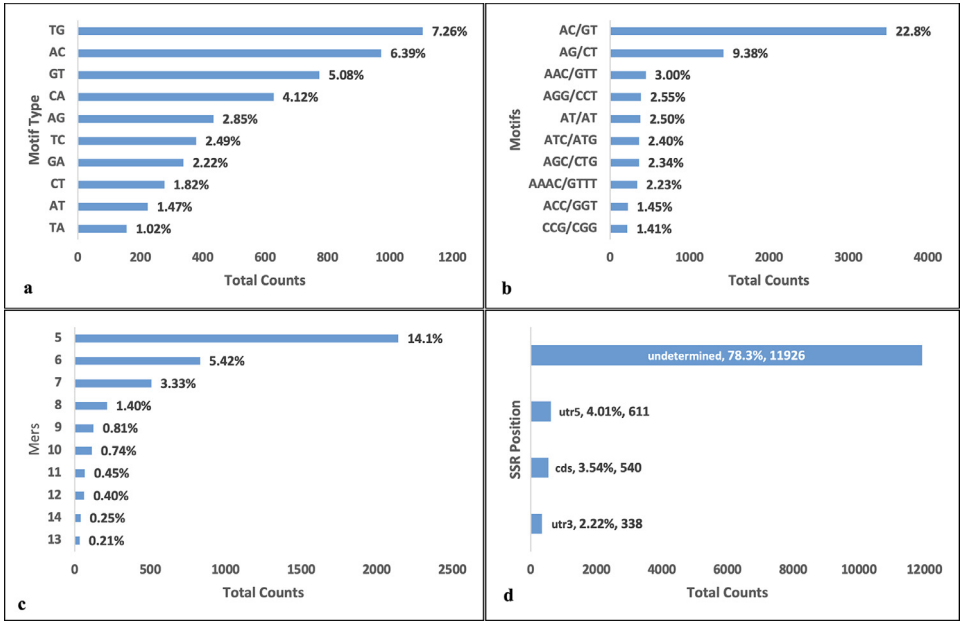

Fig. 2. Total counts and percentage of (a) SSR motif type, (b) SSR motif pairs, (c) SSR repeat numbers (d) SSR position in the transcriptome.

primer pairs are listed in Table S1. The primer size ranged between 18-23 bp with an annealing temperature starting from 57 ℃ to 60℃. GC content of the primers was between 30–70% and the final product length was100 ~ 300 bp.

The assembled transcriptome consisted of 413,890 SNPs with an overall frequency of 2.8 per kb. The predicted SNPs had 248,387 transitions (C/T-121,445; A/G-126,942) and 165,503 transversions (A/T-51,301; A/C-40,930; T/G-42,600; C/G-30,672) (Table 2). As expected, transitions were more frequent than transversions. The complete set of SNPs and indels are provided in Table S2 and S3. The counts of first, second and third position of SNPs in coding region were 122,422; 18,053; and 100,273 respectively. The total number of indels were 46,642 at a frequency of 0.31 per kb.

Table 2
Summary of SNPs and indels in the transcriptome sequences of Parvocalanus crassirostris.

Type	Count	Frequency per kb
Transition		
C/T	121445	0.82
A/G	126942	0.86
Transversion		
A/T	51301	0.35
A/C	40930	0.28
T/G	42600	0.29
C/G	30672	0.21
Total	**413890**	**2.8**
SNP position in codon		
First	122422	
Second	18053	
Third	100273	
Indels	**46642**	**0.31**

3. Experimental Design, Materials and Methods

3.1. RNA isolation

For the present investigation, live copepods (n=3) were withdrawn from a monoculture maintained (22°C;16:8 h light: dark cycle; gently aerated) in Kuwait Institute for Scientific Research (KISR) laboratories. The organisms were placed in a Bogorov chamber and visually picked for RNA isolation. Approximately 50-100 organisms were ground to a fine powder with liquid nitrogen in an autoclaved mortar and pestle. 1 ml of Trizol™ (Ambion, CA) was added and the subsequent steps of purification were followed as described in Habibi et al. [1,2]. RNA concentration was checked through the HS ssRNA assay (Qubit, Invitrogen, WA). The RNA concentrations ranged from 9.0 ng/µl-25.0 ng/ µl. RIN (RNA integrity number) could not be determined for these samples since the 28S ribosomal RNA migrates with the 18S ribosomal RNA as known for many crustacean and Arthropoda species [3–5]. Approximately 500 ng of RNA was used for RNA sequencing at Novogene, Taiwan laboratories.

3.2. Transcriptome Sequencing

The RNA isolated in the above step was used for constructing cDNA libraries through the TruSeq RNA sample prep kit (Illumina, San Deigo, CA). Briefly Oligo (dT) beads were added to selectively pick the mRNA. Fragmentation was performed and first strand cDNA synthesis initiated [6] followed by second strand cDNA synthesis. The libraries were purfied through the AMPure XP beads (Beckman Coulter Genomics, Brea, CA, USA) and indexed with Illumina indices. Libraries were quantified on a Qubit 2.0 fluorometer (Invitrogen) and diluted to 1 ng/µl. The libraries were pooled at a final concentration of 2nM and sequenced on an Illumina HiSeq 2500 in a 2 × 100 paired end chemistry. A total of 10Gb data was generated per sample. Top-up sequencing was done for COP1, COP2, and COP3 samples in order to generate the required amount of data (10Gb).

3.3. De Novo *Assembly and Filtering of microsatellites (SSRs), SNPs and Indels*

Raw reads were filtered for low quality bases and Illumina adapter and barcode sequences. Thereafter the raw sequences were treated with Corset 4.6 (-f true, default, -m 10) to remove redundancy and submitted to trinity v 2.6.6 (min Kmer Coverage =3, min glue =4). The assembled file *unigenes.fa was then uploaded to Microsatellite Identification tool (MISA) to filter SSR regions with default parameter settings (1-10 2-6 3-5 4-5 5-5 6-5) [7]. Subsequently short oligonucleotide primers were designed against the filtered SSR regions employing Primer 3 v 0.4.0 software applying the standard primer designing (primer size: 18–23 bp; annealing temperature-57–62°C; GC content-30–70%; final product size-100–400 bp) parameters [8]. Assembled transcriptome was used as a reference for variant calling. Fasta files of three generations of copepods were converted to vcf files via SAMtools v 1.9 (varFilter-Q 20 -d 1 -D 100) [9]. Single Nucleotide Polymorphisms (SNPs) and Indels were identified in the three generations of copepods [10].

Ethics Statements

Not applicable.

Data Availability

RNA Seq of Parvocalanus crassirostris (Original data) (National Centre for Biotechnology).
Transcriptome Assembly of Parvocalanus crassirostris (Original data) (National Centre for Biotechnology).

CRediT Author Statement

Nazima Habibi: Conceptualization, Writing – original draft, Writing – review & editing; **Saif Uddin:** Conceptualization, Writing – original draft, Writing – review & editing; **Montaha Behebehani:** Formal analysis, Resources, Project administration; **Mohd Wasif Khan:** Software, Data curation, Visualization; **Nasreem Abdul Razzack:** Methodology, Investigation; **Faiz Shirshikhar:** Methodology, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank Kuwait Foundation for Advancement of Sciences and Kuwait Institute for Scientific Research for funding this research

References

[1] N. Habibi, M. Behbehani, S. Uddin, F. AL Salameen, A. Shajan, F. Zakir, A safe and effective sample collection method for assessment of SARS-CoV-2 in aerosol samples, in: ALRCSMPJMVPPKF Arriola (Ed.), Environmental Resilience and Transformation in Times of COVID-19, Elsevier, 2021.
[2] N. Habibi, S. Uddin, M. Behbehani, N. Abdul Razzack, F.Z. Hussain, A. Shajan, SARS-CoV-2 in hospital air as revealed by comprehensive respiratory viral panel sequencing, Infect. Prevent. Pract. 4 (1) (2022) 100199, doi:10.1016/j.infpip.2021.100199.
[3] S. Asai, R. Sanges, C. Lauritano, P.K. Lindeque, F. Esposito, A. Ianora, et al., De novo transcriptome assembly and gene expression profiling of the copepod Calanus helgolandicus feeding on the PUA-producing diatom Skeletonema marinoi, Marine Drugs 18 (8) (2020) 392.
[4] S.D. McCarthy, M.M. Dugon, Power AM. 'Degraded'RNA profiles in Arthropoda and beyond, PeerJ 3 (2015) e1436.
[5] F. Yadetie, N.R. Brun, J. Giebichenstein, K. Dmoch, K. Hylland, K. Borgå, et al., Transcriptome responses in copepods Calanus finmarchicus, Calanus glacialis and Calanus hyperboreus exposed to phenanthrene and benzo [a] pyrene, Marine Genom. 65 (2022) 100981.
[6] E. Russo, C. Lauritano, G. d'Ippolito, A. Fontana, D. Sarno, E. von Elert, et al., RNA-Seq and differential gene expression analysis in Temora stylifera copepod females with contrasting non-feeding nauplii survival rates: an environmental transcriptomics study, BMC Genom. 21 (1) (2020) 1–22.
[7] N. Habibi, F. Al Salameen, M. Rahman, V. Kumar, S. Al Amad, A. Shajan, et al., Draft Genome Sequence and SSR mining data of Acacia pachyceras Schwartz, Data Br. (2022) 108031.
[8] N. Habibi, F. Al Salameen, N. Vyas, M.H. Rahman, V. Kumar, A. Shajan, F. Zakir, N.A. Razzack, B. Al Doaij, Genome survey and genetic characterization of Acacia pachyceras O. Schwartz, Front Plant Sci 14 (2023) 1062401.
[9] F. Al Salameen, N. Habibi, S. Al Amad, B. Al Doaij, Genetic Diversity of Rhanterium eppaposum Oliv. Populations in Kuwait as Revealed by GBS, Plants 11 (11) (2022) 1435.
[10] J. Ning, M. Wang, C. Li, S. Sun, Transcriptome sequencing and de novo analysis of the copepod Calanus sinicus using 454 GS FLX, PloS One 8 (5) (2013) e63741.