Annotated de novo transcriptomes generated from six co-occurring species of calanoid copepods from the R/V Tiglax TXF18, TXS19, TXF15, TXF17 in the Gulf of Alaska from 2015-2019

Website: https://www.bco-dmo.org/dataset/908689
Data Type: experimental, Cruise Results
Version: 1
Version Date: 2024-07-02

Project
» Collaborative Proposal: Optimizing Recruitment of Neocalanus copepods through Strategic Timing of Reproduction and Growth in the Gulf of Alaska (Neocalanus Gulf of Alaska)
» Collaborative Research: Molecular profiling of the ecophysiology of dormancy induction in calanid copepods of the Northern Gulf of Alaska LTER site (Diapause preparation)

Contributors	Affiliation	Role
Hartline, Daniel K.	University of Hawai'i at Mānoa (PBRC)	Principal Investigator
Lenz, Petra H.	University of Hawai'i at Mānoa (PBRC)	Scientist, Contact
Cieslak, Matthew C.	University of Hawai'i at Mānoa (PBRC)	Technician, Data Manager
Merchant, Lynne M.	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
The dataset includes the annotation files of nine high-quality de novo transcriptomes generated from shotgun assemblies of short-sequence reads. The species are ecologically-important members of sub-arctic North Pacific marine zooplankton communities. The de novo assemblies included one generated several years ago plus eight new ones generated from six co-occurring species of calanoid copepods in the Gulf of Alaska. The transcriptomes include the first published ones for Neocalanus plumchrus, Neocalanus cristatus, Eucalanus bungii and Metridia pacifica and three for Neocalanus flemingeri and two for Calanus marshallae. Total RNA from single individuals was used to construct gene libraries that were sequenced on an Illumina Next-Seq platform. Short-sequence reads were assembled with Trinity software and resulting transcripts were annotated using the SwissProt database with additional functional annotation using gene ontology terms and enzyme function. The annotations files are the first ones published for these species. The integrated dataset can be used for quantitative inter- and intra-species comparisons of gene expression patterns across biological processes using the annotations. These data are further described in the following publications: Hartline, et al. (2023) (DOI: 10.1038/s41597-023-02130-1), Roncalli, et al. (2022) (DOI: 10.1111/mec.16354), and Roncalli, et al. (2019) (DOI: 10.1038/s42003-019-0565-5)

Table of Contents

Coverage

Location: Gulf of Alaska
Spatial Extent: **N**:60.6667 **E**:-147.6667 **S**:59.845 **W**:-149.4667
Temporal Extent: 2015-05-10 - 2019-04-30

Dataset Description

These data are further described in the following publications:

Hartline, D. K., Cieslak, M. C., Castelfranco, A. M., Lieberman, B., Roncalli, V., & Lenz, P. H. (2023). De novo transcriptomes of six calanoid copepods (Crustacea): a resource for the discovery of novel genes. *Scientific Data*, 10(1). https://doi.org/10.1038/s41597-023-02130-1

Roncalli, V., Niestroy, J., Cieslak, M. C., Castelfranco, A. M., Hopcroft, R. R., & Lenz, P. H. (2022). Physiological acclimatization in high-latitude zooplankton. *Molecular Ecology*, 31(6), 1753–1765. Portico. https://doi.org/10.1111/mec.16354

Roncalli, V., Cieslak, M. C., Germano, M., Hopcroft, R. R., & Lenz, P. H. (2019). Regional heterogeneity impacts gene expression in the subarctic zooplankter Neocalanus flemingeri in the northern Gulf of Alaska. *Communications Biology*, 2(1). https://doi.org/10.1038/s42003-019-0565-5

Methods & Sampling

Sample collection: Zooplankton were collected from depth (2015, 2017, 2018, and 2019) at two stations in Prince William Sound: “PWS2” (Lat: 60°32′. N, Long: -147°48.2′ W) and “PWS3” (Lat: 60°40.0′ N, Long: -147°40.0′ W,) and the Gulf station “GAK1” (Lat: 59º50.7′ N, Long: -149º28′ W). Collection date, station and depth stratum for each individual are given in Hartline et al. (2023) and Roncalli et al. (2019). Zooplankton collections were made using vertical net tows with either a QuadNet with two 150 µm and two 53 µm mesh nets (April and May collections), or a multiple opening and closing plankton net (0.25 m2 cross-sectional area; 150 µm mesh nets; Multinet-Midi, Hydro-Bios; September collections). Zooplankton samples were diluted, and copepods were sorted under a dissection microscope to select individuals from the target species. Briefly, live and undamaged individuals were identified and staged using morphological criteria and preserved in RNALater Stabilization Reagent. Preserved copepods were frozen first in -20ºC during the cruises, and then transferred to −80°C until further processing. Species identification were confirmed through the COI sequence in the assembled transcriptomes.

Total RNA extraction, library construction, RNA sequencing and quality control: For each target species, total RNA was extracted from individuals using QIAGEN RNeasy Plus Mini Kit (catalog # 74134) in combination with a Qiashredder column (catalog # 79654). Selection for sequencing was based on high RNA yields and purity of extraction (RIN>8). The final list included pre-adults (CV) for *Neocalanus flemingeri* (n=3), *Neocalanus cristatus* (n=1), *Calanus marshallae* (n=2), *Eucalanus bungii* (n=1), an adult male (developmental stage CVI) for *Neocalanus plumchrus* (n=1) and an adult female for *Metridia pacifica* (n=1). Total RNA was shipped on dry ice to the Georgia Genomics Bioinformatics Core (https://dna.uga.edu) for RNA-Seq. There, double-stranded cDNA libraries (KAPA Stranded mRNA-Seq Kit, with KAPA mRNA Capture Beads (cat #KK8421]) from each individual were multiplexed and sequenced using an Illumina Next-Seq 500 instrument (High-Output Flow Cell, 150 bp, paired end). Quality of each RNA-Seq library was reviewed with the FastQC software28. From each RNA-Seq library, low quality reads were removed using FASTQ Toolkit (v. 2.2.5 within BaseSpace). Illumina adaptors, reads <50 bp long, reads with an average Phred score <30 and the first 12 bp from each read, were removed from each library. The same workflow was applied to all nine datasets.

Data Processing Description

De novo assembly, mapping, core-gene statistics: Individual de novo transcriptomes were generated from each RNA-Seq dataset at the National Center for Genome Analysis Support's (NCGAS; Indiana University, Bloomington, IN, USA) Mason Linux cluster using Trinity software (v. 2.4.0, except *N. plumchrus*, v. 2.0.6). Initial evaluation involved self-mapping of reads against the respective de novo assembly using Bowtie2 software (v. 2.3.5.1). Completeness of each de novo assembly was evaluated using Benchmarking Universal

Single-Copy Orthologs (BUSCO) software31 by searching each assembly for the presence of eukaryote “core” genes using the Arthropoda database as reference (BUSCO version 5.3.2, dataset: arthropoda_odb10 (2020-09-10, 90 genomes, 1,013 BUSCOs). RNA-Seq data and transcriptome shotgun assemblies (TSAs) have been deposited with links to BioProject accession numbers PRJNA496596, and PRJNA662858 in the NCBI (National Center for Biotechnology Information) BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/).

Functional annotation: Assemblies were functionally annotated against the NCBI Swiss-Prot protein and UniProt databases. Initial annotations were obtained by using the BLASTx algorithm on a local BLAST webserver with a Beowulf cluster using the Swiss-Prot protein database (downloaded February 2021) as reference and a threshold E-value of 10-5. Transcripts with BLAST annotations were then searched against the Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway databases using UniProt.

BCO-DMO Processing Description

Processing steps for submitted annotated individual transcriptome files and metadata

The following are the submitted annotated transcriptome files that are processed to include metadata information and then concatenated together.

Filename: n-flem_CV2015_GAK1_AccAnnots_multispp_Aug23_q-fix-2.csv
Description: Neocalanus flemingeri 2015 stage CV annotated transcriptome

Filename: Nf2018_CV_Acc&Annot-single-rev_quoted.csv
Description: Neocalanus flemingeri 2018 stage CV annotated transcriptome

Filename: Nf2019_CV-PWS2_Acc&Annot.csv
Description: Neocalanus flemingeri 2019 stage CV annotated transcriptome

Filename: Np2015_maleR1_Acc&Annot.csv
Description: Neocalanus plumchrus 2015 stage Adult Male annotated transcriptome

Filename: Nc2017_CV_Acc&Annot.csv
Description: Neocalanus cristatus 2017 stage CV annotated transcriptome

Filename: Cm2017_CV_Acc&Annot.csv
Description: Calanus marshallae 2017 stage CV annotated transcriptome

Filename: Cm2018_CV_Acc&Annot.csv
Description: Calanus marshallae 2018 stage CV annotated transcriptome

Filename: Eb2017_CV_Acc&Annot.csv
Description: Eucalanus bungii 2017 stage CV annotated transcriptome

Filename: Mp2017_AF_Acc&Annot.csv
Description: Metridia pacifica 2017 stage Adult Female annotated transcriptome

---

A metadata table with NCBI accession numbers was created using information from the submitter and by gathering accession numbers from NCBI for each BioProject.

Metadata from Submitter via email:
File Name: Summary Table for the Annotated Transcriptomes in BCO.docx
Title: Summary table from Lenz of metadata for transcriptome annotation files

Table columns
Species, Annot Filename, Stage/Collection information (which includes the Life Stage, Sex if indicated, Collection date, station, and collection depth range in meters), and NCBI TSA#

The accession number GHLB01000000 was changed to GHLB00000000 because the TSA project records accessions begin with a four-letter prefix of the TSA project followed by eight zeroes. A TSA project master accession (GenBank accession value) adds an extension of '.1' which indicates the version.

This explains it further:
See this 2013 article from NCBI on the definition of the TSA master accession number:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531190/

TSA projects will now contain a master record, in addition to records representing each of the assembled contigs. TSA will be using a similar accession number scheme to WGS as well. Like WGS accessions, the new TSA accessions have a four-letter prefix, representing the TSA project, followed by a two-digit version number and a six-digit contig number. For example, GAAA01000020 is contig 20 from the first version of TSA project GAAA. The TSA project master records have accessions that begin with the four-letter prefix followed by eight zeroes (e.g. GAAA00000000) and are indexed in the Nucleotide database.

A metadata table was created including metadata from the submitter:
File name: transcriptome_annotations_metadata_table.csv
Title: DM created metadata table of supporting information for each transcriptome annotation

Station lat and lon values from the 'Methods & Sampling' section of the submission were converted to decimal degrees with 6 digit precision.

A column named TSA_project_accession, with project accession numbers but without the version extension, was included to join the metadata table with the annotated transcriptome files later in processing, and it is removed in the final metadata table.

A column named TSA_Master_Accession which is the TSA_project_accession number with the version extension of '.1'.

From the National Center for Biotechnology Information (NCBI), various accession numbers and titles were added for each annotated transcriptome file.

Table columns:
Species, Station, Latitude, Longitude, Collection_date, Depth_range, Life_stage, Sex, BioProject, TSA_project_accession, TSA_Master_Accession, Study_Accession, Study_Title, Experiment_Accession, Experiment_Title, SRA_Accession, BioSample, Sample_Accession

---

1) Used the BCO-DMO data processor laminar, to load in the submitted annotated transcriptomes files and the DM created metadata table. The fill value '#N/A' which stands for 'no result found' was kept.

2) Created a new column for each annotated transcriptome file, TSA_project_accession, to join on the metadata table using regular expressions to extract the first 4 letters of each TSA accession number and then concatenate 8 zeros. This is the NCBI format for a TSA project accession number.

3) Joined each annotated transcriptome file with the metadata table on the key TSA_project_accession.

4) Removed the parameter TSA_project_accession, since it was only used for the Join process. The TSA project accession number with the version will be recorded as the parameter TSA_master_accession.

5) Added a suffix ‘.1’ to each TSA accession number to indicate version 1.

6) Renamed Accession# to Genbank_accession which is the TSA accession value with a version extension.

7) Renamed parameters in each annotated transcriptome file by removing commas and parentheses, and replacing spaces with underscores to follow BCO-DMO parameter naming conventions.

8) Reordered the parameters so that experiment metadata parameters are at the beginning of each annotated transcriptome file and most NCBI accession numbers are at the end

9) Using a second laminar processor pipeline, loaded in the individual annotated transcriptomes files that were created by laminar processing above and concatenated all the individual annotated transcriptome files into the primary dataset file.

10) Saved the output of the laminar processing to the supplemental files of the metadata table and the individual transcriptome annotation files.

Created a Species WoRMS taxonomy file species_list.csv by using the WoRMS website to create a table of the species name and their corresponding AphiaID and LSID values.

Checked that all the species names in the transcriptome annotation files matched those found in WoRMS.

Problem Description

For the NCBI metadata, the sample accession number SAMN16094688 and the SRP accession number SRP289633 are the same for both the 2017 and 2018 species Calanus Marshallae studies.

The submitter replied on 4/2/2024 “we need to remove the collection date for the biosample. NCBI Biosamples can have multiple sequence archives that come under the umbrella of a single Biosample description. In this specific case, we organized the biosamples by species, since the focus is on the assembly, TSA. The collection date information is found in the BCO-DMO file and in the publication of the data.”

There is a title error at NCBI for Neocalanus flemingeri 2019 with the experiment accession number SRX9453434, where the title is “Neocalanus finmarchicus” and it should be “Neocalanus flemingeri.” The submitter is working on fixing this.

[table of contents | back to top]

Data Files

File
908689_v1_annotated_transcriptomes.csv (Comma Separated Values (.csv), 611.74 MB) MD5:03f268cb5fdcc7a709a26ece6a68b58d Primary data file for dataset ID 908689, version 1 Combined file of the following supplemental annotated transcriptome files: Neocalanus flemingeri 2015 stage CV annotated transcriptome Neocalanus flemingeri 2018 stage CV annotated transcriptome Neocalanus flemingeri 2019 stage CV annotated transcriptome Neocalanus cristatus 2017 stage CV annotated transcriptome Neocalanus plumchrus 2015 stage adult male annotated transcriptome Calanus marshallae 2017 stage CV annotated transcriptome Calanus marshallae 2018 stage CV annotated transcriptome Eucalanus bungii 2017 stage CV annotated transcriptome Metridia pacifica 2017 stage adult female annotated transcriptome

[table of contents | back to top]

Supplemental Files

File
Calanus marshallae 2017 stage CV annotated transcriptome filename: supplemental_files/calanus_marshallae_stage_cv_2017_annotated_transcriptome.csv (Comma Separated Values (.csv), 66.50 MB) MD5:1657695e90292c91706d3bde9856e8de Parameter names and definitions are the same as the main combined dataset file.
Calanus marshallae 2018 stage CV annotated transcriptome filename: supplemental_files/calanus_marshallae_stage_cv_2018_annotated_transcriptome.csv (Comma Separated Values (.csv), 76.30 MB) MD5:c857228e480359010d628a54936ba774 Parameter names and definitions are the same as the main combined dataset file.

File
Eucalanus bungii 2017 stage CV annotated transcriptome filename: supplemental_files/eucalanus_bungii_stage_cv_2017_annotated_transcriptome.csv (Comma Separated Values (.csv), 34.28 MB) MD5:fb139f58502704304fe27d4a72b662c0 Parameter names and definitions are the same as the main combined dataset file.
Metadata table including NCBI accessions filename: supplemental_files/metadata_table_with_ncbi_accessions.csv (Comma Separated Values (.csv), 2.23 KB) MD5:b8d2113a435e1ec58fa0b41ae37d8f70 Metadata table including NCBI accession values for all the supplemental annotated transcriptome files. Columns are: Species, Station, Latitude, Longitude, Collection_date, Depth_range, Life_stage, Sex, BioProject, TSA_Master_Accession, Study_Accession, Study_Title, Experiment_Accession, Experiment_Title, SRA_Accession, BioSample, Sample_Accession The parameter definitions are the same as the main dataset file and are described in the parameters section.
Metridia pacifica 2017 stage adult female annotated transcriptome filename: supplemental_files/metridia_pacifica_stage_adult_female_2017_annotated_transcriptome.csv (Comma Separated Values (.csv), 103.88 MB) MD5:7c0038b596bee11d3256bb931d08e769 Parameter names and definitions are the same as the main combined dataset file.
Neocalanus cristatus 2017 stage CV annotated transcriptome filename: supplemental_files/neocalanus_cristatus_stage_cv_2017_annotated_transcriptome.csv (Comma Separated Values (.csv), 80.95 MB) MD5:b0ea2f052f4e7685e81ccad05c23086f Parameter names and definitions are the same as the main combined dataset file.
Neocalanus flemingeri 2015 stage CV annotated transcriptome filename: supplemental_files/neocalanus_flemingeri_stage_cv_2015_annotated_transcriptome.csv (Comma Separated Values (.csv), 49.92 MB) MD5:f7ab9b5de56d3d0482bc899ddd1bd4a1 Parameter names and definitions are the same as the main combined dataset file.
Neocalanus flemingeri 2018 stage CV annotated transcriptome filename: supplemental_files/neocalanus_flemingeri_stage_cv_2018_annotated_transcriptome.csv (Comma Separated Values (.csv), 75.87 MB) MD5:4efc489558896f89d2d6859ad36e5486 Parameter names and definitions are the same as the main combined dataset file.
Neocalanus flemingeri 2019 stage CV annotated transcriptome filename: supplemental_files/neocalanus_flemingeri_stage_cv_2019_annotated_transcriptome.csv (Comma Separated Values (.csv), 60.02 MB) MD5:a42cfb6560b0bb3ddb62122b1244937f Parameter names and definitions are the same as the main combined dataset file.
Neocalanus plumchrus 2015 stage adult male annotated transcriptome filename: supplemental_files/neocalanus_plumchrus_stage_adult_male_2015_annotated_transcriptome.csv (Comma Separated Values (.csv), 64.02 MB) MD5:4eabeb7f7cb0f0d3c013129a413cf4f2 Parameter names and definitions are the same as the main combined dataset file.
Species WoRMS taxonomy filename: species_list.csv (Comma Separated Values (.csv), 889 bytes) MD5:92d2ec05ffd87f4cf786844631e2187f Taxonomic identifiers AphiaID and LSID for the species of the annotated transcriptomes

[table of contents | back to top]

Related Publications

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Software

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E.,

Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. Nature Genetics, 25(1), 25–29. https://doi.org/10.1038/75556
Methods

BaseSpace Labs. (n.d.). FASTQ Toolkit (Version 2.2.5) [Computer software]. Illumina. https://www.illumina.com/products/by-type/informatics-products/basespace-sequence-hub/apps/fastq-toolkit.html
Software

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Agivetova, R., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bursteinas, B., Bye-A-Jee, H., Coetzee, R., Cukura, A., Da Silva, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., … Teodoro, D. (2020). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research, 49(D1), D480–D489. https://doi.org/10.1093/nar/gkaa1100
Methods

Boeckmann, B. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Research, 31(1), 365–370. https://doi.org/10.1093/nar/gkg095
Methods

Gene Ontology Consortium; Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, Ebert D, Feuermann M, Gaudet P, Harris NL, Hill DP, Lee R, Mi H, Moxon S, Mungall CJ, Muruganugan A, Mushayahama T, Sternberg PW, Thomas PD, Van Auken K, Ramsey J, Siegele DA, Chisholm RL, Fey P, Aspromonte MC, Nugnes MV, Quaglia F, Tosatto S, Giglio M, Nadendla S, Antonazzo G, Attrill H, Dos Santos G, Marygold S, Strelets V, Tabone CJ, Thurmond J, Zhou P, Ahmed SH, Asanitthong P, Luna Buitrago D, Erdol MN, Gage MC, Ali Kadhum M, Li KYC, Long M, Michalak A, Pesala A, Pritazahra A, Saverimuttu SCC, Su R, Thurlow KE, Lovering RC, Logie C, Oliferenko S, Blake J, Christie K, Corbani L, Dolan ME, Drabkin HJ, Hill DP, Ni L, Sitnikov D, Smith C, Cuzick A, Seager J, Cooper L, Elser J, Jaiswal P, Gupta P, Jaiswal P, Naithani S, Lera-Ramirez M, Rutherford K, Wood V, De Pons JL, Dwinell MR, Hayman GT, Kaldunski ML, Kwitek AE, Laulederkind SJF, Tutaj MA, Vedi M, Wang SJ, D'Eustachio P, Aimo L, Axelsen K, Bridge A, Hyka-Nouspikel N, Morgat A, Aleksander SA, Cherry JM, Engel SR, Karra K, Miyasato SR, Nash RS, Skrzypek MS, Weng S, Wong ED, Bakker E, Berardini TZ, Reiser L, Auchincloss A, Axelsen K, Argoud-Puy G, Blatter MC, Boutet E, Breuza L, Bridge A, Casals-Casas C, Coudert E, Estreicher A, Livia Famiglietti M, Feuermann M, Gos A, Gruaz-Gumowski N, Hulo C, Hyka-Nouspikel N, Jungo F, Le Mercier P, Lieberherr D, Masson P, Morgat A, Pedruzzi I, Pourcel L, Poux S, Rivoire C, Sundaram S, Bateman A, Bowler-Barnett E, Bye-A-Jee H, Denny P, Ignatchenko A, Ishtiaq R, Lock A, Lussi Y, Magrane M, Martin MJ, Orchard S, Raposo P, Speretta E, Tyagi N, Warner K, Zaru R, Diehl AD, Lee R, Chan J, Diamantakis S, Raciti D, Zarowiecki M, Fisher M, James-Zorn C, Ponferrada V, Zorn A, Ramachandran S, Ruzicka L, Westerfield M. The Gene Ontology knowledgebase in 2023. Genetics. 2023 May 4;224(1):iyad031. doi: 10.1093/genetics/iyad031. PMID: 36866529; PMCID: PMC10158837.
Methods

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology, 29(7), 644–652. doi:10.1038/nbt.1883
Software

Hartline, D. K., Cieslak, M. C., Castelfranco, A. M., Lieberman, B., Roncalli, V., & Lenz, P. H. (2023). De novo transcriptomes of six calanoid copepods (Crustacea): a resource for the discovery of novel genes. Scientific Data, 10(1). https://doi.org/10.1038/s41597-023-02130-1
Results

Kanehisa, M. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research, 28(1), 27–30. doi:10.1093/nar/28.1.27
Methods

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Science, 28(11), 1947–1951. Portico. https://doi.org/10.1002/pro.3715
Methods

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., & Ishiguro-Watanabe, M. (2022). KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Research, 51(D1), D587–D592. https://doi.org/10.1093/nar/gkac963
Methods

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9(4), 357–359. doi:10.1038/nmeth.1923
Software

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and

Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution, 38(10), 4647–4654. https://doi.org/10.1093/molbev/msab199
Software

Roncalli, V., Cieslak, M. C., Germano, M., Hopcroft, R. R., & Lenz, P. H. (2019). Regional heterogeneity impacts gene expression in the subarctic zooplankter Neocalanus flemingeri in the northern Gulf of Alaska. Communications Biology, 2(1). https://doi.org/10.1038/s42003-019-0565-5
Results

Roncalli, V., Niestroy, J., Cieslak, M. C., Castelfranco, A. M., Hopcroft, R. R., & Lenz, P. H. (2022). Physiological acclimatization in high-latitude zooplankton. Molecular Ecology, 31(6), 1753–1765. Portico. https://doi.org/10.1111/mec.16354
Results

[table of contents | back to top]

Related Datasets

IsRelatedTo

Lenz, P. H., Cieslak, M. C., Roncalli, V., Hartline, D. K. (2025) **Molecular identification of genetic variants of Neocalanus flemingeri in the Gulf of Alaska from samples collected from 2015 to 2023.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2025-02-21 doi:10.26008/1912/bco-dmo.954181.1 [view at BCO-DMO]
Relationship Description: This dataset (908689) was cited in the methods of another dataset (954181) "...The reference consisted of consensus sequences obtained by comparing sequences downloaded from NCBI and from de novo assemblies (Hartline et al., 2023; Hartline et al., 2024 [BCO-DMO dataset 908689]). "

Lenz, P. H., Hartline, D. K., Roncalli, V., Block, L. N., Niestroy, J. L., Cieslak, M. C. (2024) **Gene expression profiles for Neocalanus flemingeri pre adults (CV) exposed to four different experimental food conditions collected from the M/V Dora in the Gulf of Alaska at station GAK1 from April 2019.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2024-05-30 doi:10.26008/1912/bco-dmo.914459.1 [view at BCO-DMO]

University of Hawaii at Manoa (2018). Neocalanus flemingeri, Neocalanus flemingeri pre adult (CV). 2018/10. NCBI:BioProject: PRJNA496596 [Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA496596.

University of Hawaii at Manoa (2018). Neocalanus plumchrus, Neocalanus cristatus, Calanus marshallae, Eucalanus bungii, Metridia pacifica, Ocean ZOOP-Zooplankton transcriptomes. 2020/09. NCBI:BioProject: PRJNA662858.[Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA662858

[table of contents | back to top]

Parameters

Parameter	Description	Units
seq_id	The Trinity software names assembled by Trinity software are named hierarchically grouping sequences by similarity	unitless

Genbank_accession	NCBI (National Center for Biotechnology Information) Accession number for nucleotide sequence in the Transcriptome Shotgun Assembly (TSA) Sequence Database. The accession number is a unique identifier assigned to a record in the sequence database GenBank at NCBI. It is of the format [alphabetical prefix][series of digits].[version]. A change in the record is tracked by an integer extension of the accession number, an Accession.version identifier. The initial version of a sequence has the extension ".1".	unitless
Species	A taxonomic binomial that consists of a genus name followed by the species name	unitless
Station	Station identifier	unitless
Latitude	Sampling location latitude, south is negative	decimal degrees
Longitude	Sampling location longitude, west is negative	decimal degrees
Collection_date	Collection date of organism	unitless
Depth_range	Collection depth range	meters (m)
Life_stage	Organism life history stage	unitless
Sex	sex	unitless
Entry	Uniprot KB entry identifies the top BLAST (Basic Local Alignment Search Tool) hit sequence to the assembled nucleotide sequence, searches were conducted using the blastx algorithm, #N/A indicates that there was no positive BLAST hit that met the e-value threshold	unitless
Entry_name	Uniprot entry name for the top hit, #N/A = no hit	unitless
evalue	Expected probability, e-value is the number of expected hits of similar quality (score) that could be found just by chance, cut-off value used for the annotation: E-value = 10-5, #N/A = no hit	unitless
Protein_names	Protein names, #N/A = no hit	unitless
Gene_names	Gene name based on the reference genome of the top hit, #N/A = no hit	unitless

Organism	Species name of top hit sequence, #N/A = no hit	unitless
Cross_reference_KEGG	KEGG identification based on the cross-reference of protein annotation and the Kyoto Encyclopedia of Genes and Genomes database, #N/A = no hit	unitless
Gene_ontology_IDs	GO terms based on cross-reference of protein annotation to the Gene Ontology Resource for functional identification as to Biological Process (BP), Cellular Component (CC) and Molecular Function (MF), proteins are typically involved in multiple processes, #N/A = no hit	unitless
Gene_ontology_GO	Identification of GO term with a functional description, #N/A = no hit	unitless
Gene_ontology_biological_process	Gene ontology terms and descriptions associated with Biological Process, #N/A = no hit	unitless
Gene_ontology_cellular_component	Gene ontology terms and descriptions associated with Cellular Component, #N/A = no hit	unitless
Gene_ontology_molecular_function	Gene ontology terms and descriptions associated with Molecular Function, #N/A = no hit	unitless
BioProject	NCBI BioProject accession	unitless
TSA_Master_Accession	NCBI (National Center for Biotechnology Information) Accession number for the Transcriptome Shotgun Assembly (TSA) master record. The accession number is a unique identifier assigned to a master record in the database GenBank at NCBI. It is of the format [alphabetical prefix][00000000].[version]. A change in the record is tracked by an integer extension of the accession number, an Accession.version identifier. The initial version of a sequence has the extension ".1".	unitless
Study_Accession	NCBI SRA study accession (SRP)	unitless
Study_Title	Title of NCBI SRA study	unitless
Experiement_Accession	Experiment metadata table (SRX) accession number in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA).	unitless

Experiment_Title	Title of the the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) experiment metadata table (SRX)	unitless
SRA_Accession	Run accession in the Sequence Read Archive (SRA) at NCBI	unitless
BioSample	NCBI BioSample accession	unitless
Sample_Accession	NCBI SRA Sample accession (SRS)	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina Next-Seq 500
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	Desktop sequencer
Generic Instrument Description	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

Dataset-specific Instrument Name	Dissection microscope
Generic Instrument Name	Microscope - Optical
Dataset-specific Description	An optical microscope variant
Generic Instrument Description	Instruments that generate enlarged images of samples using the phenomena of reflection and absorption of visible light. Includes conventional and inverted instruments. Also called a "light microscope".

Dataset-specific Instrument Name	Multinet-Midi, Hydro-Bios
Generic Instrument Name	MultiNet
Dataset-specific Description	Hydro-Bios Multinet-Midi with a 0.25 m2 cross-sectional area and 150 µm mesh
Generic Instrument Description	The MultiNet© Multiple Plankton Sampler is designed as a sampling system for horizontal and vertical collections in successive water layers. Equipped with 5 or 9 net bags, the MultiNet© can be delivered in 3 sizes (apertures) : Mini (0.125 m2), Midi (0.25 m2) and Maxi (0.5 m2). The system consists of a shipboard Deck Command Unit and a stainless steel frame to which 5 (or 9) net bags are attached by means of zippers to canvas. The net bags are opened and closed by means of an arrangement of levers that are triggered by a battery powered Motor Unit. The commands for actuation of the net bags are given via single or multi-conductor cable between the Underwater Unit and the Deck Command Unit. Although horizontal collections typically use a mesh size of 300 microns, mesh sizes from 100 to 500 may also be used. Vertical collections are also common. The shipboard Deck Command Unit displays all relevant system data, including the actual operating depth of the net system.

Dataset-specific Instrument Name	QuadNet
Generic Instrument Name	Plankton Net
Dataset-specific Description	Two 150 µm and two 53 µm mesh nets
Generic Instrument Description	A Plankton Net is a generic term for a sampling net that is used to collect plankton. It is used only when detailed instrument documentation is not available.

[table of contents | back to top]

Deployments

TXF18

Website	https://www.bco-dmo.org/deployment/910684
Platform	R/V Tiglax
Report	https://nga.lternet.edu/wp-content/uploads/2019/04/Cruise-Report-TXF18.pdf
Start Date	2018-09-11
End Date	2018-09-25
Description	NGA LTER Fall cruise

TXS19

Website	https://www.bco-dmo.org/deployment/910688
Platform	R/V Tiglax
Report	https://nga.lternet.edu/wp-content/uploads/2019/10/Cruise-Report-TXS19.pdf
Start Date	2019-04-26
End Date	2019-05-08
Description	NGA LTER Summer cruise

TXF15

Website	https://www.bco-dmo.org/deployment/852877
Platform	R/V Tiglax
Report	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Coastal/seward.html
Start Date	2015-09-09
End Date	2015-09-21
Description	Latitude North boundary (decimal degrees): 60.5298 Latitude South boundary (decimal degrees): 57.7747 Longitude West Boundary (decimal degrees): -149.4755 Longitude East Boundary (decimal degrees): -147.5105

TXF17

Website	https://www.bco-dmo.org/deployment/852883
Platform	R/V Tiglax
Report	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Coastal/seward.html
Start Date	2017-09-09
End Date	2017-09-22
Description	Latitude North boundary (decimal degrees): 60.6753 Latitude South boundary (decimal degrees): 57.7923 Longitude West Boundary (decimal degrees): . -149.4853 Longitude East Boundary (decimal degrees): -147.503

TXS15

Website	https://www.bco-dmo.org/deployment/917221
Platform	R/V Tiglax
Report	https://www.ncei.noaa.gov/access/ocean-carbon-acidification-data-system/oceans/Coastal/seward.html
Start Date	2015-05-05
End Date	2015-05-11

[table of contents | back to top]

Project Information

Collaborative Proposal: Optimizing Recruitment of Neocalanus copepods through Strategic

Timing of Reproduction and Growth in the Gulf of Alaska (Neocalanus Gulf of Alaska)

Coverage: Gulf of Alaska; Seward Line

NSF abstract:

The Gulf of Alaska supports a diverse and productive marine community that includes many commercially important fishes. Toward the base of this food web are small planktonic crustaceans that serve as the primary food source for many of these fish, as well as seabirds and marine mammals. The copepod Neocalanus flemingeri is one of these crustaceans, and it experiences rapid population growth during each spring's algal, or phytoplankton, bloom. An apparent mismatch between the presence of the youngest stages of the copepod, or nauplii, in early winter and the unpredictable timing of the spring phytoplankton bloom several months later raises important questions about when females reproduce and how this relates to survival and growth of nauplii. Two types of dormancy, diapause in adult females and physiological quiescence in nauplii, may be the key to the success of this copepod species. Timing and duration of the egg-laying period by adult females is linked to emergence from diapause. In addition, nauplii may enter a state of physiological quiescence while food resources are low, resuming growth after phytoplankton levels increase. This research will address a long-standing goal of biological oceanographers to understand dormancy and its role in controlling population cycles in marine copepods. It will use new technologies in molecular biology called transcriptomics to catalog the messages used by the cells to control copepod life processes, in this case those related to dormancy in adults and nauplii. Undergraduate students and a postdoctoral investigator will be trained in interdisciplinary research, and students from Native Hawaiian and Native Alaskan groups will be targeted for participation. Fishing is a major industry in the Gulf of Alaska, and outreach will focus on communicating the role copepods play in marine ecosystems. New content, including images, will be generated for existing websites: the Seward Line long-term observation program, the Alaska Ocean Observing System and the Gulf Watch Alaska Program.

Recruitment to the Neocalanus flemingeri spring population is dependent on successful emergence from diapause followed by reproduction, survival, and growth of the next generation. Individual-based models have made significant progress in predicting population growth in calanoid copepods using food, temperature, and advection as key environmental factors. Few of these models include predictors for naupliar recruitment, however, because little is known about this part of the life cycle given sampling difficulties and the lack of biomarkers to evaluate physiological state. This study will leverage existing monitoring efforts to track the N. flemingeri population during the winter and early spring. The research team will combine laboratory and field approaches to determine duration and synchronization of reproduction in emerging females and strategies for naupliar survival during low food conditions. Zooplankton samples will be processed to enumerate nauplii to species and to determine physiological condition of both nauplii and adult females. Gene expression studies will develop molecular markers for female dormancy and reproductive readiness and for naupliar growth and possible dormancy, which in turn will be used to evaluate field collected individuals. This will be the first comprehensive study to combine molecular and traditional tools to connect diapausing adults, naupliar production, and the resulting spring population of copepodites.

Collaborative Research: Molecular profiling of the ecophysiology of dormancy induction in calanid copepods of the Northern Gulf of Alaska LTER site (Diapause preparation)

Coverage: Northern Gulf of Alaska LTER

NSF Award Abstract:
The sub-arctic Pacific sustains major fisheries with nearly all commercially important species depending either directly or indirectly on lipid-rich copepods (Neocalanus flemingeri, Neocalanus plumchrus, Neocalanus cristatus and Calanus marshallae). In turn, these species depend on a short-lived spring algal bloom for growth and the accumulation of lipid stores in order to complete an annual life cycle that includes a period of dormancy. The intellectual thrust of this project measures how the timing and magnitude of algal blooms affect preparation for dormancy using a combination of field and experimental observations. The Northern Gulf of Alaska - with four calanid species that experience dormancy, steep environmental gradients, well-described phytoplankton bloom dynamics, and a concurrent NSF-LTER program - provides an unusual opportunity to identify the factors that affect dormancy preparation. Education and outreach plans are integrated with the

research. Educational efforts focus on interdisciplinary opportunities for undergraduate, graduate and post-doctoral trainees. The project will generate content for existing graduate and undergraduate courses. U. of Alaska Fairbanks and U. Hawaii at Manoa are Alaska Native and Native Hawaiian Serving Institutions, and students from these groups will be recruited to participate in the project. Because fishing is a major industry in the Gulf of Alaska, outreach will communicate the role copepods play in marine ecosystems using the concept of a dynamic food web tied to production cycles.

Diapause (dormancy) and the accompanying accumulation of lipids in copepods have been identified as key drivers in high latitude ecosystems that support economically important fisheries, including those of the Gulf of Alaska. While the disappearance of lipid-rich copepods has been linked to severe declines in fish stocks, little is known about the environmental conditions that are required for the successful completion of the copepod's life cycle. A physiological profiling approach that measures relative gene expression will be used to test two alternative hypotheses: the lipid accumulation window hypothesis, which holds that individuals enter diapause only after they have accumulated sufficient lipid stores, and the developmental program hypothesis, which holds that once the diapause program is activated, progression occurs independent of lipid accumulation. The specific objectives are: 1) determine the effect of food levels during N. flemingeri copepodite stages on progression towards diapause using multiple physiological and developmental markers; 2) characterize the seasonal changes in the physiological profile of N. flemingeri across environmental gradients and across years; 3) compare physiological profiles across co-occurring calanid species (N. flemingeri, Neocalanus plumchrus, Neocalanus cristatus and Calanus marshallae); and 4) estimate the reproductive potential of the overwintering populations of N. flemingeri. The broader scientific significance includes the acquisition of new genomic data and molecular resources that will be made publicly available through established data repositories, and the development of new tools for routinely obtaining physiological profiles of copepods.

This award reflects NSF's statutory mission and has been deemed worthy of support through evaluation using the Foundation's intellectual merit and broader impacts review criteria.

NOTE: Petra Lenz is a former Principal Investigator (PI) and Andrew Christie is a former Co-Principal Investigator (Co-PI) on this project (award #1756767). Daniel Hartline is the PI listed for the award #1756767 and is now a former Co-PI on this project.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-1459235
NSF Division of Ocean Sciences (NSF OCE)	OCE-1756767
NSF Division of Ocean Sciences (NSF OCE)	OCE-1756859

[table of contents | back to top]