Full genome and transcriptome sequence assembly of the non-model organism Kellet's whelk, Kelletia kelletii

Website: https://www.bco-dmo.org/dataset/945292
Data Type: Other Field Results
Version: 1
Version Date: 2024-12-04

Project
» Collaborative Research: RUI: Combined spatial and temporal analyses of population connectivity during a northern range expansion (KW connectivity)

Contributors	Affiliation	Role
White, Crow	California Polytechnic State University San Luis Obispo (Cal Poly (SLO))	Principal Investigator, Contact
Christie, Mark	Purdue University	Co-Principal Investigator
Davidson, Jean	California Polytechnic State University San Luis Obispo (Cal Poly (SLO))	Co-Principal Investigator
Toonen, Robert J.	University of Hawai'i at Mānoa (HIMB)	Co-Principal Investigator
Anderson, Paul	California Polytechnic State University San Luis Obispo (Cal Poly (SLO))	Scientist
Daniels, Benjamin	California Polytechnic State University San Luis Obispo (Cal Poly (SLO))	Scientist
Lee, Andy	Purdue University	Scientist
López, Cataixa	University of Hawai'i at Mānoa (HIMB)	Scientist
Soenen, Karen	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
Understanding the genomic characteristics of non-model organisms can bridge research gaps between ecology and evolution. However, the lack of a reference genome and transcriptome for these species makes their study challenging. Here, we complete the first full genome and transcriptome sequence assembly of the non-model organism Kellet's whelk, Kelletia kelletii, a marine gastropod exhibiting a poleward range expansion coincident with climate change. We used a combination of Oxford Nanopore Technologies, PacBio, and Illumina sequencing platforms and integrated a set of bioinformatic pipelines to create the most complete and contiguous genome documented among the Buccinoidea superfamily to date. Genome validation revealed relatively high completeness with low missing metazoan Benchmarking Universal Single-Copy Orthologs (BUSCO) and an average coverage of ~70x for all contigs. Genome annotation identified a large number of protein-coding genes similar to some other closely related species, suggesting the presence of a complex genome structure. Transcriptome assembly and analysis of individuals during their period of peak embryonic development revealed highly expressed genes associated with specific Gene Ontology (GO) terms and metabolic pathways, most notably lipid, carbohydrate, glycan, and phospholipid metabolism. We also identified numerous heat shock proteins (HSPs) in the transcriptome and genome that may be related to coping with thermal stress during the sessile life history stage. A robust reference genome and transcriptome for the non-model organism K. kelletii provide resources to enhance our understanding of its ecology and evolution and potential mechanisms of range expansion for marine species facing environmental changes.

Table of Contents

Coverage

Location: Cal Poly Pier, Avila Beach, CA 93424, USA (35.169800, -120.740800)
Spatial Extent: **N**:35.17 **E**:-119.57 **S**:34.25 **W**:-120.75
Temporal Extent: 2019 - 2020

Dataset Description

Description of linked resources for this dataset, all links can be found in the related dataset section.

- All codes and parameters used for the bioinformatic analyses carried out are available at https://github.com/bndaniel/Kellets-whelk-genome-assembly and are archived at Zenodo, doi:10.5281/zenodo.13274364
- The Kelletia kelletii genome and transcriptome produced by this study have been deposited in Dryad, doi:10.5061/dryad.w0vt4b8zn
- All raw sequence data, including the PacBio sequel 2, Nanopore MinION, and Illumina NovaSeq DNA sequencing, as well as the Illumina NovaSeq RNA sequencing, are deposited in the NCBI Sequence Read Archive (SRA) under PRJNA999368 and PRJNA1000198
- VCF files for the all-SNPs and DEG-SNPs data sets are deposited in Dryad, doi:10.5061/dryad.qbzkh18s3
- The scripts used in this project are hosted in the public repository https://github.com/ChristieLab/kellets_whelk_rnaseq and archived at Zenodo, doi:10.5281/zenodo.14187737

Methods & Sampling

gDNA was extracted using HMW Circulomics Standard TissueRuptor Protocol. DNA was further cleaned and concentrated using the DNeasy PowerClean Pro Cleanup Kit. Libraries were prepared using the PacBio HiFi SMARTbell from Ultra-low DNA input procedures. Libraries were prepared for sequencing using the Binding kit 2.2.

RNA was extracted using Trizol and RNA phase separation. After fragmentation, the first strand cDNA was synthesized using random hexamer primers, followed by the second strand cDNA synthesis using dTTP for non-directional library preparation. Messenger RNA was purified from total RNA using poly-T oligo-attached magnetic beads. Libraries were prepared using end repair, A-tailing, adapter ligation, size selection, amplification, and purification.

Detailed methods in Daniels et al., (2023), see related publications

BCO-DMO Processing Description

* Merged RNA and DNA sequence files
* Split lat_lon column into latitude and longitude column

[table of contents | back to top]

Data Files

File
945292_v1_kelletia.csv(Comma Separated Values (.csv), 46.23 KB) MD5:7d0b036e1424a609a44b6bc904eedda5 Primary data file for dataset ID 945292, version 1

[table of contents | back to top]

Related Publications

Daniels, B. N., Andrasz, C. L., Zarate, N., Lee, A., López, C., Anderson, P., Toonen, R. J., Christie, M. R., White, C., & Davidson, J. M. (2023). De novo genome and transcriptome assembly of Kelletia kelletii, a coastal gastropod and fisheries species exhibiting a northern range expansion. Frontiers in Marine Science, 10. https://doi.org/10.3389/fmars.2023.1278131
Results

Lee, A., Daniels, B. N., Hemstrom, W., López, C., Kagaya, Y., Kihara, D., Davidson, J. M., Toonen, R. J., White, C., & Christie, M. R. (2024). Genetic adaptation despite high gene flow in a range-expanding population. Molecular Ecology. Portico. https://doi.org/10.1111/mec.17511
Results

[table of contents | back to top]

Related Datasets

IsRelatedTo

Ben Daniels, & Cassidy Andrasz. (2024). *bndaniel/Kellets-whelk-genome-assembly: Kellet's whelk Genome and Transcriptome assembly and analysis* (Version v1.0.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.13274364 https://doi.org/10.5281/zenodo.13274364

Cal Poly. Genome sequencing of Kellet's whelk. 2023/07. In: BioProject [Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; 2011-. Available from: http://www.ncbi.nlm.nih.gov/bioproject/PRJNA999368. NCBI:BioProject: PRJNA999368.

California Polytechnic State University. Total mRNA sequencing of Kellet's whelk embryos. 2023/07. In: BioProject [Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information; 2011-. Available from: http://www.ncbi.nlm.nih.gov/bioproject/PRJNA1000198. NCBI:BioProject: PRJNA1000198.

Daniels, B. (2023). *Kellet's whelk genome and transcriptome assembly* (Version 5) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.W0VT4B8ZN https://doi.org/10.5061/dryad.w0vt4b8zn

Lee, A. (2024). *Kellet's Whelk Common Garden RNA-Seq Analyses*. Zenodo. https://doi.org/10.5281/ZENODO.14187737 https://doi.org/10.5281/zenodo.14187737

Lee, A., Daniels, B., López, C., Davidson, J., Toonen, R., White, C., & Christie, M. (2024). *SNPs derived from a common garden experiment across the biogeographic range of Kelletia kelletii* (Version 2) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.QBZKH18S3 https://doi.org/10.5061/dryad.qbzkh18s3

[table of contents | back to top]

Parameters

Parameter	Description	Units
Run	NCBI SRA run accession ID	unitless

Assay_Type	Type of sequencing assay performed: WGS or RNA-Seq	unitless
AvgSpotLen	The average length of the spots (reads) in the run	unitless
Bases	The total number of bases sequenced	unitless
BioProject	NCBI Bioproject accession ID	unitless
BioSample	NCBI Biosample accession ID	unitless
BioSampleModel	NCBI BioSample package representing type of biosample and required attributes	unitless
Bytes	Filesize in bytes for datastore file	unitless
Center_Name	The name of the sequencing center	unitless
Collection_Date	Collection date of organism	unitless
Consent	Information on the consent for data usage	unitless
DATASTORE_filetype	The file type stored in the NCBI DataStore (e.g., FASTQ, BAM)	unitless
DATASTORE_provider	The provider of the DataStore where files are kept	unitless
DATASTORE_region	The geographical region of the DataStore	unitless
Ecotype	A population within a given species displaying genetically based, phenotypic traits that reflect adaptation to a local habitat, e.g., Columbia	unitless
env_broad_scale	Broad-scale environmental context	unitless
env_local_scale	Local-scale environmental context	unitless
env_medium	Material displaced by the entity at time of sampling	unitless
Experiment	The NCBI identifier for the sequencing experiment	unitless

geo_loc_name_country	Geographic location of the origin of the sample: country = USA	unitless
geo_loc_name_country_continent	Geographic location of the origin of the sample: continent = North America	unitless
geo_loc_name	Geographic location of the origin of the sample	unitless
Instrument	The sequencing instrument used (e.g., Illumina MiSeq)	unitless
isol_growth_condt	Isolation and growth condition specifications	unitless
latitude	Latitude of sampling location, south is negative	decimal degrees
longitude	Longitude of sampling location, west is negative	decimal degrees
Library_Name	Unique identifier for the sequencing library (can be the sample name repeated)	unitless
LibraryLayout	Single or paired end sequencing reads	unitless
LibrarySelection	Selection used for sequencing library	unitless
LibrarySource	Source material of sequencing library	unitless
Organism	Name of organism from which the sample was taken: Kelletia kelletii	unitless
Platform	Sequencing platform manufacturer	unitless
ReleaseDate	Date of release of NCBI data to the public	unitless
create_date	Date of creation of NCBI submission	unitless
version	The version of the NCBI submission: 1	unitless
Sample_Name	Unique sample name	unitless
source_material_id	Unique identifier assigned to the sequenced material	unitless

SRA_Study	NCBI SRA study accession ID	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	ILLUMINA
Generic Instrument Name	Automated DNA Sequencer
Generic Instrument Description	General term for a laboratory instrument used for deciphering the order of bases in a strand of DNA. Sanger sequencers detect fluorescence from different dyes that are used to identify the A, C, G, and T extension reactions. Contemporary or Pyrosequencer methods are based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step.

Dataset-specific Instrument Name	Nanopore MinION
Generic Instrument Name	Automated DNA Sequencer
Generic Instrument Description	General term for a laboratory instrument used for deciphering the order of bases in a strand of DNA. Sanger sequencers detect fluorescence from different dyes that are used to identify the A, C, G, and T extension reactions. Contemporary or Pyrosequencer methods are based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step.

Dataset-specific Instrument Name	PacBio sequel 2
Generic Instrument Name	Automated DNA Sequencer
Generic Instrument Description	General term for a laboratory instrument used for deciphering the order of bases in a strand of DNA. Sanger sequencers detect fluorescence from different dyes that are used to identify the A, C, G, and T extension reactions. Contemporary or Pyrosequencer methods are based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step.

[table of contents | back to top]

Project Information

Collaborative Research: RUI: Combined spatial and temporal analyses of population connectivity during a northern range expansion (KW connectivity)

Coverage: California, USA and Baja, Mexico coast

NSF Award Abstract:
Where do young marine fish and shellfish come from? This project aims to improve our understanding of how coastal marine populations are connected in space and time. Coastal populations are replenished through the arrival of minuscule larvae that have been dispersed for weeks to months in the open ocean after spawning at remote sites. The combination of the long dispersal period of marine fish and shellfish larvae and the varying ocean currents results in complex patterns of "connectivity" among populations near and far. Identifying these patterns of connectivity is fundamental to marine science and critical for effective fisheries management and conservation, yet it remains an unresolved component of marine ecology. The study species is currently expanding its biogeographic range up the U.S. west coast. By genetically analyzing individuals from across the species' range, including offspring spawned in the laboratory by experimentally-crossed individuals collected in the field from throughout the species historical and expanded range, certain genes can serve to differentiate populations along the coast. The team leverages the statistical power of these geographically-informative genes to assign thousands of young collected in the field to the source populations that spawned them (across the species' range and over multiple years). The team then quantifies patterns of connectivity over multiple years, and tests fundamental hypotheses on the spatial scale, temporal variability, biogeographic patterns, and biophysical drivers of population connectivity. The project trains approximately two dozen U.S. university students in molecular ecology and marine science, as well as creating intellectual linkages among Ph.D.-granting and non-Ph.D.-granting universities. The project also supports further development of a K-12 education program that uses SCUBA diving and videography to teach elementary school students Next Generation Science Standards and train them for careers in science, technology, engineering and mathematics.

Using a kelp forest gastropod and fisheries species (Kellet's whelk, Kelletia kelletii), this project combines genome-wide Restriction site Associated DNA (RAD) loci with transcriptomic loci identified from common-garden laboratory crosses of individuals from the species' historical and expanded range to identify geographically-informative loci that maximize power for individual assignment testing. Leveraging the combined power of these loci, genetic assignment of approximately three thousand recruit samples to 20 putative source populations allows the team to construct three independent years of connectivity matrices and test some of the most fundamental questions in marine ecology, including: 1) Are marine populations open or closed and at what scales? 2) To what degree is the evolutionary pattern of gene flow represented by single versus multiple generations of connectivity events? And, 3) How spatially heterogeneous and temporally variable is population connectivity? Can one year of connectivity data predict anything about the next? Additionally, by focusing on a range-expanding species with common life history traits, the team addresses a number of questions with broad applicability and significant ecological and societal implications: 4) How much is population connectivity influenced by post-recruitment demographic and evolutionary processes? 5) How well-connected are historic- and expanded-range populations? And, of particular relevance to climate change, 6) Are El Nino oceanographic conditions, which are predicted to increase in frequency and intensity this century, driving the poleward range expansion of this coastal marine species? By coupling common-garden experimental crosses to identify maximally-informative transcriptomic loci with genomic RAD analysis of field samples, this project aims to accurately and precisely quantify marine population connectivity in high gene flow species with large population sizes.

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.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-1924537

[table of contents | back to top]