Pooled sequencing data of field-collected Eurytemora affinis copepods from nine locations in North America collected October 2012 to March 2014

Website: https://www.bco-dmo.org/dataset/883227
Data Type: Other Field Results
Version: 1
Version Date: 2022-11-12

Project
» Evolutionary Responses to Global Changes in Salinity and Temperature (Evolutionary genomics of a copepod)

Contributors	Affiliation	Role
Lee, Carol E.	University of Wisconsin (UW-Madison)	Principal Investigator
Stern, David B.	University of Wisconsin (UW-Madison)	Scientist
Gerlach, Dana Stuart	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
This dataset consists of pooled whole-genome sequencing data (Pool-seq) of Eurytemora affinis complex copepods collected from nine locations across North America. Each sample consists of 100 pooled individuals with an approximate 1:1 sex ratio, sequenced using the Nextera DNA library preparation kit and Illumina HiSeq platform. Population sampling targeted nine wild populations of the E. affinis complex from two genetically distinct lineages (the Gulf and Atlantic clades) including both native saline and invaded freshwater locations. The data were generated to investigate evolutionary genomic factors underlying successful invasion of freshwater habitats and were collected by Drs. Martin Bontrager, David B. Stern, and Carol E. Lee of the University of Wisconsin-Madison.

Table of Contents

Coverage

Spatial Extent: **N**:48.002 **E**:-69.423 **S**:29.254 **W**:-96.797
Temporal Extent: 2012 - 2014

Methods & Sampling

Copepods were collected by Drs. Martin Bontrager, David B. Stern, and Carol E. Lee of the University of Wisconsin-Madison using plankton tows from nine locations across North America:

- Braddock Bay, Lake Ontario, NY (43.307, -77.706)
- Milwaukee, Lake Michigan, WI (43.051, -87.882)
- Montmagny, St. Lawrence Estuary, QC (46.99, -70.55)
- L'Isle Verte, St. Lawrence Estuary, QC (48.002, -69.423)

- Lake Texoma, Red. River, OK (33.882, -96.797)
- Lake Eufaula, Arkansas River, OK (35.146, -95.627)
- Louisville, Ohio River, KY (38.26, -85.747)
- Cocodrie Bayou, Gulf of Mexico, LA (29.254, -90.664)
- Taylor Bayou, Gulf of Mexico, TX (29.883, -94.051)

Population sampling targeted nine wild populations of the *Eurytemora affinis* complex (copepods) from two genetically distinct lineages (the Gulf and Atlantic clades) including both native saline and invaded freshwater locations. Salinity measurements from each sampling site were collected using a handheld refractometer.

Plankton was brought back to the lab and sorted. *Eurytemora affinis* complex copepods were pooled into samples consisting of 100 individuals with an approximate 1:1 sex ratio, and subjected to DNA extraction. DNA sequencing data were collected using the Nextera DNA library preparation kit (Illumina, Inc., San Diego, CA, USA). Libraries were sequenced on an Illumina HiSeq platform at the University of Maryland, School of Medicine, Institute for Genome Sciences. The data were generated to investigate evolutionary genomic factors underlying successful invasion of freshwater habitats. Patterns of genetic diversity within and among populations were analyzed to detect signatures of natural selection.

Additional methods and results can be found in Stern & Lee (2020).
Scripts to process allele frequency files used in the PoolSeq analyses available in Stern (2022).
Sequence data is archived under NCBI BioProject PRJNA610547 (see Related Datasets below)

Data Processing Description

Raw reads were trimmed and filtered of adapter sequences, low-complexity sequences and low-quality (Q<15) bases using BBDuk in the BBTools package (Bushnell, 2014). **Processed read**s were mapped to the repeat-masked *E. affinis* complex (Atlantic clade) draft reference genome using BWA-MEM v.0.7.17 (Li, 2013). **Paired-end read**s that did not align concordantly with BWA-MEM were aligned as single-end reads using NextGenMap v.0.5.5 (Sedlazeck et al., 2013) to aid in the alignment of diverged sequences. The combined read-mapping procedure achieved a mean mapping rate of 95.09 ± 1.42%. Duplicate reads were removed using Picard v.2.18.27 (Soifer, 2022) and regions around insertions or deletions were realigned using GATK v.3.8 (Van der Auwera & O'Connor, 2020). SAMtools v.1.3.1(Li et al., 2009) was used to convert BAM files into mpileup format after removing low-quality alignments and bases (Q<20). Sites within 3 bp (base pair) of an insertion or deletion were removed and the filtered mpileup was converted to sync format using PoPoolation2 (Kofler, Pandey, Schlotterer, 2011). The R package poolfstat v.1.0 (Gautier et al., 2021) was used to detect bi-allelic SNPs with a global minor allele frequency >0.05, at least four reads were required for a base call, and a minimum of 20 and a maximum of 200 total read counts were required for all populations. In total, 6,635,765 SNPs passed these filters when considering all nine populations. A total of 7,565,621 and 5,323,780 SNPs were called for the Atlantic and Gulf clades, respectively.

Software
(See also Related Publications section below)

- BBTools v38.33 (Bushnell, 2014)
- BWA-MEM v0.7.17 (Li, 2013)
- NextGenMap v0.5.5 (Sedlazeck, 2013)
- Picard v2.18.27 (Soifer, 2022)
- GATK v3.8 (Van der Auwera & O'Connor, 2020)
- SAMtools v1.3.1 (Li et al., 2009)
- PoPoolation v1.2.2 (Kofler et al., 2011)
- PoPoolation2 v1.013 (Kofler, Pandey, Schlotterer, 2011)
- poolfstat v.1.0 (Gautier et al., 2021)
- TreeMix v1.13 (Pickrell & Pritchard, 2012)
- phytools v.0.6 (Revell, 2011)
- BayeScan 2, BayeScan 3 (Foll & Gaggiotti, 2008)
- BayPass v2.1 (Gautier, 2015)
- BetaScan2 (Siewert & Voight, 2020)
- GenMap (Pockrandt et al., 2020)
- edgeR (Robinson et al, 2010)
- BedTools v2.28 (Quinlan & Hall, 2010)
- BEDOPS (Neph et al., 2012)

- Gowinda (Kofler & Schlotterer, 2012)
- custom code (Stern, 2022 for https://github.com/TheDBStern/poolseq_utils)

BCO-DMO Processing
- Converted date to Y-M-D format
- Separated latitude and longitude into separate columns and converted to decimal degrees
- Added a column for BioProject
- Added dates of collection based on Supplemental File information
- Sorted by Collection_Date

[table of contents | back to top]

Data Files

File
pooled_sequence.csv (Comma Separated Values (.csv), 1.52 KB) MD5:49100700996893c40ca155f9302e8355 Primary data file for dataset ID 883227

[table of contents | back to top]

Supplemental Files

File
Copepod sampling dates and locations filename: Copepod_sampling_dates_locations.pdf (Portable Document Format (.pdf), 26.09 KB) MD5:8e8166ef7edf9de0ae86600e091af9d0 Copepod sampling dates and locations

[table of contents | back to top]

Related Publications

Bushnell, B. (2014). BBTools software package. http://bbtools.jgi.doe.gov
Software

,
Software

Foll, M., & Gaggiotti, O. (2008). A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective. Genetics, 180(2), 977–993. https://doi.org/10.1534/genetics.108.092221 https://doi.org/DOI:10.1534/genetics.108.092221
Software

Gautier, M. (2015). Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics, 201(4), 1555–1579. https://doi.org/10.1534/genetics.115.181453
Software

Gautier, M., Vitalis, R., Flori, L., & Estoup, A. (2021). f-Statistics estimation and admixture graph construction with Pool-Seq or allele count data using the R package poolfstat. Molecular Ecology Resources, 22(4), 1394–1416. Portico. https://doi.org/10.1111/1755-0998.13557
Software

Kofler, R., & Schlotterer, C. (2012). Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. Bioinformatics, 28(15), 2084–2085. https://doi.org/10.1093/bioinformatics/bts315
Software

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., & Schlötterer, C. (2011). PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. PLoS ONE, 6(1), e15925. https://doi.org/10.1371/journal.pone.0015925

Software

Kofler, R., Pandey, R. V., & Schlotterer, C. (2011). PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics, 27(24), 3435–3436. https://doi.org/10.1093/bioinformatics/btr589
Software

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (Version 2). arXiv. https://doi.org/10.48550/ARXIV.1303.3997 https://doi.org/10.48550/arXiv.1303.3997
Software

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., … Homer, N. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. doi:10.1093/bioinformatics/btp352
Methods
,
Software

Neph, S., Kuehn, M. S., Reynolds, A. P., Haugen, E., Thurman, R. E., Johnson, A. K., Rynes, E., Maurano, M. T., Vierstra, J., Thomas, S., Sandstrom, R., Humbert, R., & Stamatoyannopoulos, J. A. (2012). BEDOPS: high-performance genomic feature operations. Bioinformatics, 28(14), 1919–1920. https://doi.org/10.1093/bioinformatics/bts277
Software

Pickrell, J. K., & Pritchard, J. K. (2012). Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLoS Genetics, 8(11), e1002967. https://doi.org/10.1371/journal.pgen.1002967
Software

Pockrandt, C., Alzamel, M., Iliopoulos, C. S., & Reinert, K. (2020). GenMap: ultra-fast computation of genome mappability. Bioinformatics, 36(12), 3687–3692. https://doi.org/10.1093/bioinformatics/btaa222
Software

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033
Software

Revell, L. J. (2011). phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution, 3(2), 217–223. Portico. https://doi.org/10.1111/j.2041-210x.2011.00169.x
Software

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2009). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616
Software

Sedlazeck, F. J., Rescheneder, P., & von Haeseler, A. (2013). NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics, 29(21), 2790–2791. https://doi.org/10.1093/bioinformatics/btt468
Software

Siewert, K. M., & Voight, B. F. (2020). BetaScan2: Standardized Statistics to Detect Balancing Selection Utilizing Substitution Data. Genome Biology and Evolution, 12(2), 3873–3877. https://doi.org/10.1093/gbe/evaa013
Software

Soifer I (2022) Picard GitHub. http://broadinstitute.github.io/picard/
Software

Stern, D. B. (2022). TheDBStern/poolseq_utils: First release(Version v0.0.1) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.7300110 https://doi.org/10.5281/zenodo.7300110
Methods

Stern, D. B., & Lee, C. E. (2020). Evolutionary origins of genomic adaptations in an invasive copepod. Nature Ecology & Evolution, 4(8), 1084–1094. https://doi.org/10.1038/s41559-020-1201-y
Results

Van der Auwera, G.A & O'Connor, B.D. (2020). Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition). O'Reilly Media. https://isbnsearch.org/isbn/9781491975190
Software

[table of contents | back to top]

Related Datasets

IsSupplementedBy

University of Wisconsin - Madison. Eurytemora affinis pooled whole-genome sequencing. 2020/03. 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/PRJNA610547. NCBI:BioProject: PRJNA610547. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA610547

[table of contents | back to top]

Parameters

Parameter	Description	Units
BioProject	NCBI BioProject identifier	unitless
Collection_Date	Date of sample collection	unitless
Latitude	Latitude of copepod sample collection	decimal degrees
Longitude	Longitude of copepod sample collection (West is negative)	decimal degrees
geo_loc_name	Country and location of copepod sample collection	unitless
Salinity	Water salinity of collected sample	PSU
Origin	Copepod origin indicating whether the population was native or invasive	unitless
Run	NCBI SRA run number	unitless
Bases	Total sequenced DNA bases	unitless
BioSample	NCBI BioSample number	unitless
Experiment	NCBI SRA experiment number	unitless
Library_Name	Sample code	unitless
Instrument	Illumina sequencing instrument used to generate the data	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina HiSeq 2000
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	Libraries were sequenced on an Illumina HiSeq platform at the University of Maryland, School of Medicine, Institute for Genome Sciences.
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	
Generic Instrument Name	Plankton Net
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.

Dataset-specific Instrument Name	handheld refractometer
Generic Instrument Name	Refractometer
Dataset-specific Description	Salinity measurements from each sampling site were collected using a handheld refractometer.
Generic Instrument Description	A refractometer is a laboratory or field device for the measurement of an index of refraction (refractometry). The index of refraction is calculated from Snell's law and can be calculated from the composition of the material using the Gladstone-Dale relation. In optics the refractive index (or index of refraction) n of a substance (optical medium) is a dimensionless number that describes how light, or any other radiation, propagates through that medium.

[table of contents | back to top]

Project Information

Evolutionary Responses to Global Changes in Salinity and Temperature (Evolutionary genomics of a copepod)

Coverage: St. Lawrence estuary, Gulf of Mexico, Great Lakes, Baltic Sea

NSF Award Abstract:

Drastic changes in the global water cycle and increases in ice melt are causing the freshening of Northern coastal seas. The combination of both reduced salinity and increased temperature will likely act in concert to reduce populations of estuarine and marine organisms. Data indicate that reduced salinity and high temperature would each increase the energy costs as well as reduce survival and reproduction of the common copepod Eurytemora affinis. This project will examine the joint effects of salinity reduction and temperature increase on the evolutionary responses of populations of E. affinis in the wild, as well as in selection experiments in the laboratory. This study will provide novel insights into responses of organisms to climate change, as no study has analyzed the joint impacts of salinity and temperature on evolutionary responses, and relatively few studies have examined the impacts of declining salinity. In general, how selection acts at the whole genome level is not well understood, particularly for non-model organisms. As a dominant estuarine copepod, E. affinis is among the most important species sustaining coastal food webs and fisheries in the Northern Hemisphere, such as salmon, herring, and anchovy. Thus, insights into its evolutionary responses with changing climate have important implications for sustainability of fisheries and food security. Two graduate students from historically underrepresented groups will be trained during this project. The project will have additional societal benefits, including development of educational modules for K-12 students and international collaboration.

This study will address the following questions: (1) To what extent could populations evolve in response to salinity and temperature change, and what are the fitness and physiological costs? (2) How will populations respond to the impacts of salinity-temperature interactions? (3) Do wild populations show evidence of natural selection in response to salinity and temperature? To analyze the evolutionary responses of E. affinis

populations to the coupled impacts of salinity and temperature, the investigator will perform laboratory selection experiments and population genomic surveys of wild populations. Selection experiments constitute powerful tools for determining the rate, trajectory, and limits of adaptation. During laboratory selection, evolutionary shifts in fitness-related traits and genomic expression will be examined, as well as genomic signatures of selection in response to low salinity and high temperature selection regimes. The investigator will also conduct population genomic sequencing of E. affinis populations that reside along salinity and temperature gradients in the St. Lawrence and Baltic Sea, and identify genes that show signatures of selection. The project will determine whether the loci that show signatures of selection in the wild populations are the same as those favored during laboratory selection. This reproducibility will provide greater confidence that the genes involved in adaptation to salinity and/or temperature have been captured.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-1658517

[table of contents | back to top]