

RADseq SNP genotypes of *Perkinsus marinus* and *Crassostrea virginica* from the Gulf of Mexico and eastern Atlantic shorelines collected from 2011-2019

Website: <https://www.bco-dmo.org/dataset/998962>

Data Type: Other Field Results

Version: 1

Version Date: 2026-05-19

Project

» [The genetic legacy of an Asian oyster introduction and its disease-causing parasite](#) (Oyster historical genetics)

Contributors	Affiliation	Role
Sotka, Erik	College of Charleston (CofC)	Principal Investigator
York, Amber D.	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract

For host-parasite interactions, genotype-by-sequencing can allow the simultaneous examination of host and parasite genomes and can yield insight into co-evolutionary processes. The eastern oyster, *Crassostrea virginica*, is among the most important aquacultured species in the United States. Natural and farmed oyster populations can be heavily impacted by 'dermo' disease caused by an alveolate protist, *Perkinsus marinus*. Here, we used restricted site-associated DNA sequencing (RADseq) to simultaneously examine the spatial population genetic structure of host and parasite. We analyzed 393 single-nucleotide polymorphisms (SNPs) for *P. marinus* and 52,100 SNPs for *C. virginica* from 36 individual oysters from the Gulf of Mexico (GOM) and mid-Atlantic coastline. This dataset includes metadata, methods, links to published code and processed data used to generate figures for results publication Weatherup et al. (2024, doi:10.1017/S0031182024000611) titled "Co-phylogeographic structure in a disease-causing parasite and its oyster host." This dataset also includes genetic accession identifiers for sequence data contributed to the National Center for Biotechnology Information (NCBI)'s Sequence Read Archive (SRA), available under BioProject PRJNA1074361.

Table of Contents

- [Coverage](#)
- [Dataset Description](#)
 - [Methods & Sampling](#)
 - [Data Processing Description](#)
 - [BCO-DMO Processing Description](#)
 - [Problem Description](#)
- [Related Publications](#)
- [Related Datasets](#)
- [Parameters](#)
- [Instruments](#)
- [Project Information](#)
- [Funding](#)

Coverage

Location: Gulf of Mexico and eastern Atlantic coastline of the United States; intertidal oyster beds.

Spatial Extent: N:37.81 E:-75.66 S:29.24 W:-90

Temporal Extent: 2011-10-06 - 2019-10-24

Methods & Sampling

For each oyster sampled, animals were cleaned, measured and shucked to expose soft tissues. Gill and mantle were extracted from each oyster and preserved in 95% ethanol for genetic analysis. DNA was extracted using

a QIAGEN QIAamp DNA Mini Kit. Genomic DNA concentration was measured with a Nanodrop spectrophotometer to ensure the concentration was $\geq 100 \text{ ng } \mu\text{L}^{-1}$. All positive samples in microcentrifuge tubes were randomly mixed before transport in case there was an error in future sequencing steps that could cause a loss in representation of an entire population. Once the microcentrifuge tubes containing the extracted DNA were thoroughly mixed, 20 μL of each of the 64 samples were used in subsequent library construction. The DNA of each individual oyster was digested separately with 2 restriction enzymes, EcoRI and MseI. The digested DNA fragments were then ligated to Illumina adaptors at the MseI end and with Illumina adaptors coupled with an 8–10 bp unique barcode to the EcoRI end to allow identification of the individual *in silico*. The restriction-ligation products were then PCR-amplified in 2 separate reactions using standard Illumina primers. The final PCR products were pooled and shipped for sequencing.

Fifty-nine *P. marinus*-positive samples were sequenced in 2020 at the Tufts University Genomic Services (200–400 bp fraction; single-end sequencing on an Illumina HiSeq 2500 using SE125), 42 samples were sequenced in 2022 at the University of Texas Genomic Sequencing and Analysis Facility (300–450 bp fraction; single-end sequencing on an Illumina NovaSeq SP using SR100) and 37 of each of those samples were sequenced in both runs.

Organism identifiers (Life Science Identifier (LSID)):
Crassostrea virginica (urn:lsid:marinespecies.org:taxname:140657)
Perkinsus marinus (urn:lsid:marinespecies.org:taxname:562957)

Data Processing Description

Reads were aligned to the *P. marinus* genome (GCA_000006405.1) and the *C. virginica* genome (GCF_002022765.2). The median number of reads aligned to the *P. marinus* genome (0.13 M; a range of 0.10–1.01 M) was 0.37% of the median number aligned to the *C. virginica* genome (34 M; a range of 3.52–66.89 M). This yielded a ratio of 265 *C. virginica* to *P. marinus* reads per sample. Of the total number of reads per sample, a median of 0.80 and 91.30% of reads were *P. marinus* and *C. virginica*, respectively.

After alignment of 64 samples to the *P. marinus* genome, we kept 35 samples that had between 100 K and 1 M *P. marinus* reads. Another sample identified as 6837–16, from VA had 1.1 M reads, from which we randomly downsampled 500 K reads using samtools view.

For *P. marinus*, we found all single-nucleotide polymorphisms, or SNPs [set minor allele frequency (MAF) threshold at 0] and then a custom script to find SNPs that had at least 1 read across 50% of individuals. This yielded 772 SNPs. An analysis of allele frequencies indicated that 393 SNPs were polymorphic between or within samples. Thus, all subsequent analyses are based on phred-scale genotype likelihoods of 393 polymorphic SNPs at 36 individuals. For *C. virginica*, we created genotypes from the same 36 individuals, and included SNPs with MAF > 1% and at least 1 read per sample, yielding genotypes at 52 100 SNPs.

Organism name, Life Science Identifier (LSID):
Crassostrea virginica, urn:lsid:marinespecies.org:taxname:140657
Perkinsus marinus, urn:lsid:marinespecies.org:taxname:562957

BCO-DMO Processing Description

- Loaded CSV file "meta_forBCODMO.csv" into table "998962_v1_perkinsus-crassostrea-co-phylogeny"; empty strings and "NA" treated as missing values
- Set types for 18 columns: Accession, BioProject, BioSample, geo_loc_name, host, isolate, isolation_source, organism, sample_name, siteName, State, Storage, tissue, Region, Region2 as string; Lat, Lon as number; date_collection as string
- Updated field metadata (descriptions, standard name IDs, supplied units) for all 18 columns, including NCBI accession identifiers (Accession, BioProject, BioSample), geographic coordinates (Lat, Lon), collection date, organism (Crassostrea virginica and Perkinsus marinus), tissue type, storage method, and location fields
- Converted date_collection from format "%m/%d/%y" to ISO 8601 string format "%Y-%m-%d"
- Set date_collection type to date with format "%Y-%m-%d"
- Renamed column Accession to Accession (correcting spelling)
- Output final table as "998962_v1_perkinsus-crassostrea-co-phylogeny.csv"

Problem Description

NA

[[table of contents](#) | [back to top](#)]

Related Publications

Weatherup, E. F., Carnegie, R., Strand, A. E., & Sotka, E. E. (2024). Co-phylogeographic structure in a disease-causing parasite and its oyster host. *Parasitology*, 151(7), 671–678.

<https://doi.org/10.1017/s0031182024000611> <https://doi.org/10.1017/S0031182024000611>

Results

[[table of contents](#) | [back to top](#)]

Related Datasets

Software

Erik Sotka. (2024). *esotka/WeatherupPerkinsus: Published dataset and code* (Version 1.0) [Computer software]. Zenodo. <https://doi.org/10.5281/ZENODO.13381489>

IsRelatedTo

College of Charleston (2024). Co-phylogeographic structure in a disease-causing parasite and its oyster host. 2024/02. NCBI:BioProject: PRJNA1074361. In: BioProject [Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information. Available from: <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA1074361> <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1074361>

[[table of contents](#) | [back to top](#)]

Parameters

Parameter	Description	Units
BioSample	National Center for Biotechnology Information (NCBI) BioSample accession.	unitless
sample_name	sample ID	unitless
organism	'Crassostrea virginica' LSID(urn:lsid:marinespecies.org:taxname:140657), or 'Perkinsus marinus' LSID(urn:lsid:marinespecies.org:taxname:562957)	unitless
isolate	sample ID v2	unitless
host	parasite or none (missing data identifier will vary by file type accessed. Blank in csv files).	unitless
isolation_source	Estuary	unitless
geo_loc_name	Country:State	unitless
tissue	gill/mantle or Culture	unitless
date_collection	Collection Date	unitless
siteName	SiteName long	unitless
State	US State	unitless
Region	Bay or Eastern Shore (VA)	unitless
Region2	Atlantic or Gulf	unitless
Storage	DNA stored in FFPE or Fresh	unitless
Lat	latitude	decimal degrees
Lon	longitude	decimal degrees
Accesion	National Center for Biotechnology Information (NCBI) Run accession.	unitless
BioProject	National Center for Biotechnology Information (NCBI) BioProject identifier.	unitless

Instruments

Dataset-specific Instrument Name	Illumina Sequencing machine (HiSeq 2500)
Generic Instrument Name	Automated DNA Sequencer
Generic Instrument Description	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

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

Dataset-specific Instrument Name	384-well thermocycler (Eppendorf)
Generic Instrument Name	Thermal Cycler
Generic Instrument Description	A thermal cycler or "thermocycler" is a general term for a type of laboratory apparatus, commonly used for performing polymerase chain reaction (PCR), that is capable of repeatedly altering and maintaining specific temperatures for defined periods of time. The device has a thermal block with holes where tubes with the PCR reaction mixtures can be inserted. The cycler then raises and lowers the temperature of the block in discrete, pre-programmed steps. They can also be used to facilitate other temperature-sensitive reactions, including restriction enzyme digestion or rapid diagnostics. (adapted from http://serc.carleton.edu/microbelife/research_methods/genomics/pcr.html)

Project Information

The genetic legacy of an Asian oyster introduction and its disease-causing parasite (Oyster historical genetics)

Coverage: Global

NSF abstract:

During the 20th century, the Pacific oyster *Crassostrea gigas* was deliberately introduced from its native range of coastal Asia to the estuaries of six continents. While the introduced Pacific oysters are widely aquacultured and thus can generate local economic wealth, they sometimes outcompete native oysters, and can carry microbial, animal and plant hitchhikers that negatively impact local economies and the ecological functioning of local estuaries. This study comprehensively assesses the pathways and sources of Pacific oyster introductions using a worldwide, population genetic survey. Simultaneously, the study also assesses the pathways and source of one hitchhiking protist (*Haplosporidium nelsoni*) that causes the disease MSX (multinucleated sphere X) in the Virginia oyster (*Crassostrea virginica*) along the eastern seaboard of the United States. One goal of this research is to generate management strategies that combat the negative impacts of the Pacific oyster and

its associated invaders, and minimize future invasions. A second goal is to minimize some uncertainty about the population biology of the devastating Haplosporidium parasite, and thus, increase confidence of policy makers who are managing shellfish health, restoration and commerce. By quantifying the pathways and sources of *C. gigas*, this project may inform strategies to combat negative impacts of *C. gigas* and its associated invaders, as well as minimize future invasions. Moreover, quantifying dispersal within and among populations of *H. nelsoni* along the US East Coast will provide perspective on the effectiveness of regional biosecurity measures in preventing the ongoing dispersal of this destructive pathogen via aquaculture. In addition, the project lends itself well to programs that foster critical thinking and research experience among both undergraduate and K-12 students. The project provides opportunities for 6-9 undergraduates to perform research, includes a 2-day workshop on bioinformatics for the wider undergraduate community, and facilitates ongoing opportunities for K-12 students to participate in citizen-science research.

There is a wealth of information on the source, pathways and vectors of *C. gigas* based largely on historical documents but no study has comprehensively tested whether these historical accounts are correct using a worldwide, population genetic survey. Using >14K single-nucleotide polymorphisms (SNPs) from 41 populations across five continents a high level of spatial genetic differentiation was found within the native range and differences in source populations among non-native regions. Preliminary genetic data indicated that the parasitic protist, *Haplosporidium nelsoni* arrived with *C. gigas* imports to the US Atlantic coastline and then infected the native *C. virginica*, however the native source populations, the pathways and vector from which *H. nelsoni* arrived remain unknown. This project couples high-throughput sequencing technologies and Approximate Bayesian Computing (ABC)-based models to answer the following: What are the population genomic patterns among *C. gigas* from native and non-native regions? What are the population genomic patterns of *Haplosporidium nelsoni* among Asian and North American *Crassostrea gigas* and eastern North American *C. virginica*? What were the source populations and invasion pathways of *C. gigas* and *H. nelsoni*? Identifying source locations, pathways and vectors of introduction of *C. gigas* will provide researchers with a null-model of invasion history for dozens of other non-native species that were transported with *C. gigas*. Currently, there are no verified 'vector maps' for historical shipments of *C. gigas* that are similar to those generated from modern-day or historical shipping records.

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-1924599

[[table of contents](#) | [back to top](#)]