

Sequence metadata for all sampled oysters from an experimental oyster reef restoration in Ninigret Pond, Rhode Island (USA) in 2017-2020

Website: <https://www.bco-dmo.org/dataset/1003894>

Data Type: Other Field Results

Version: 1

Version Date: 2026-08-07

Project

» [CAREER: Linking genetic diversity, population density, and disease prevalence in seagrass and oyster ecosystems](#) (Seagrass and Oyster Ecosystems)

Contributors	Affiliation	Role
Hughes, A. Randall	Northeastern University	Principal Investigator
Truskey, Sarit	Northeastern University	Co-Principal Investigator
Mickle, Audrey	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract

This dataset contains sequencing metadata for all 803 eastern oysters (*Crassostrea virginica*, urn:lsid:marinespecies.org:taxname:140657) sampled for genetic analysis and retained after SNP quality filtering as part of a multi-year oyster reef restoration experiment in Ninigret Pond, Rhode Island, USA. Oysters were sourced from four commercial hatcheries along the U.S. Atlantic coast (Maine, Massachusetts, New York, and Virginia) and sampled from hatchery source collections pre-experiment and from experimental restored reefs over time. Metadata are drawn from the NCBI Sequence Read Archive (BioProject PRJNA1280068), where the associated raw fastq files are publicly archived, and include sample identifiers, collection details, sequencing instrument/platform information, and SRA accession numbers linking each individual to its raw sequencing data. The primary genetic cluster assignment for each individual broadly correspond to the four commercial hatchery sources as described in the related individual genetic assignments dataset. A flag column indicates individuals excluded from downstream reef-level analyses for the associated manuscript, including those from reefs determined to have started as single-source reefs in fall 2018 and a small subset of individuals sampled outside the primary fall 2018 and fall 2020 time points.

Table of Contents

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

Coverage

Location: Ninigret Pond, Charlestown, Rhode Island

Spatial Extent: Lat:41.3549 Lon:-71.6929

Temporal Extent: 2017-11 - 2020-10

Methods & Sampling

These data were published in Truskey et al. 2025 (Evolutionary Applications). All figure numbers and

supplementary materials mentioned refer to Truskey et al. 2025 (Evolutionary Applications).

Restoration experiment and sample collection

In April 2017, we collaborated with local oyster farmers to source oyster eyed-larvae from four commercial hatcheries for the construction of new experimental oyster reefs in Ninigret Pond (Charlestown, Rhode Island, USA). Hatchery sources were selected across a broad geographic sampling range to maximize diversity (Figure 1a of Truskey et al. 2025). They included two regional hatcheries, one from Massachusetts (MA) and one from New York (NY), and two more distant hatcheries, one from Maine (ME) and one from Virginia (VA). A local Rhode Island hatchery facility received oyster eyed-larvae from each source hatchery, set the eyed-larvae on dead oyster and/or clam shell (hereafter referred to as spat-on-shell oysters), and distributed the spat-on-shell oysters to three local oyster growers (Grower 1: MA; Grower 2: ME, NY, VA; Grower 3: NY, VA) who maintained these juvenile oysters on separate leased oyster farm plots prior to reef construction. Immediately prior to reef construction, we collected 20 juvenile spat-on-shell oysters from each grower-hatchery source combination for genetic analysis ($n = 6$ combinations; Grower 1-MA; Grower 2-ME; Grower 2-NY; Grower 2-VA; Grower 3-NY; Grower 3-VA).

In October and November 2017, we constructed 16 subtidal reefs within a no-harvest Shellfish Management Area in Ninigret Pond. Reefs were created across four 0.025-acre experimental blocks, with four reefs per block (Figure 2 of Truskey et al. 2025). Each reef consisted of a base layer of 0.25 cubic yards of dead oyster or clam shell deployed in October and topped with 1.25 cubic yards of spat-on-shell oysters in early November from the stock grown out by the local oyster growers. Our experiment was originally designed to seed three reefs per block with a different single hatchery source from the four available hatchery sources and to seed the fourth reef per block with a mixed combination of the three sources (e.g., Reef 1: ME, Reef 2: MA, Reef 3: NY, Reef 4: ME+MA+NY). However, analysis of the hatchery samples from oyster growers collected prior to reef construction revealed an early, unintended mixing of some of the sources (Appendix S1 of Truskey et al. 2025; Figure S1 of Truskey et al. 2025), resulting in mixtures of multiple sources on 12 out of the 16 constructed reefs. Thus, we focused our analyses on the actual genetic composition of each reef as determined by genetic sampling at two time points (fall 2018, fall 2020).

In fall 2018, we haphazardly sampled live oysters from each reef on scuba or snorkel ($N = 512$ individuals total, 32 per reef). Oysters were put on ice and transported to the Northeastern University Marine Science Center where they were held at -80°C until DNA extraction. By the fall of 2020, live oyster densities on all experimental reefs had declined, reflecting mortality of the original planted oysters and a lack of recruitment, consistent with other data from this system (Barrett et al. 2024). To assess whether this mortality was associated with a consistent change in the genetic composition of surviving oysters on reefs, we repeated our sampling in fall 2020 and compared the resulting reef genetic profiles to those from fall 2018. Reef sample sizes varied at this time point due to low live abundances ($N = 249$ individuals total, ranging from 8 to 32 per reef).

DNA extraction, RADseq library preparation, and bioinformatics

Genomic DNA was extracted using the E-Z 96 Tissue DNA Kit (Omega-Biotek, Norcross, GA) following the animal tissue protocol with tissue centrifugation. Double-digest restriction-site-associated DNA (ddRAD) libraries with individually barcoded samples were prepared in three batches following Parchman et al. (2012): (1) initial hatchery samples ($n = 120$), (2) fall 2018 reef samples ($n = 480$; 30 oysters per reef across 16 reefs), and (3) fall 2020 reef samples ($n = 248$). For additional details on ddRAD library preparation, see Appendix S2 of Truskey et al. 2025. All libraries were sequenced with 100-bp single-end reads on an Illumina platform. The initial and fall 2020 batches were sequenced on a single lane of the Illumina HiSeq 2500 at Tufts University Core Facility Genomics; the fall 2018 batch was sequenced on two lanes of the Illumina NovaSeq 6000 at the University of Texas at Austin Genomic Sequencing and Analysis Facility.

Data Processing Description

SNP calling, filtering, and defining datasets

Raw sequence quality was assessed with FastQC v0.11.9 (Andrews 2010), and files were demultiplexed using the `process_radtags` function in STACKS v2.41 (Catchen et al. 2013). Read trimming, mapping, SNP calling, and genotyping were performed with the dDocent pipeline v2.9.4 (Puritz et al. 2014), with reads mapped to the *C. virginica* genome pruned for haplotigs (Puritz et al. 2024). Variant calling was performed using FreeBayes v1.3.6 (Garrison and Marth 2012). SNP loci were then filtered using vcftools v0.1.16 (Danecek et al. 2011) following standard iterative quality filtering procedures for RADseq-generated SNP data (O'Leary et al. 2018),

including filters for minor allele count, base quality, genotype call rate, and read depth, followed by individual-level filtering for missing data. Because genetic assignment approaches can be sensitive to the SNP dataset used, we generated 12 SNP sets varying in minor allele frequency (MAF) threshold, locus-level filtering for missingness specific to the fall 2020 library, and pruning for high linkage disequilibrium (LD), to evaluate the robustness of downstream genetic cluster assignments to these bioinformatic decisions. Filtering outcomes for each SNP set are detailed in Table S1 of Truskey et al. 2025.

The individuals presented in this dataset are all of those retained after SNP filtering steps and used in subsequent analyses to generate individual assignments to genetic clusters.

BCO-DMO Processing Description

- Loaded biosample_parsed.csv with missing values defined as empty string, "nd", and "NA"
- Loaded SraRunTable_updated_BCODMO_resubmission.csv as table "sra_run_metadata", with missing values defined as empty string, "nd", and "NA"
- Split the lat_lon column (e.g. "41.354865 N, -71.692945 W") into four intermediate columns: lat_value, lat_hemisphere, lon_value, lon_hemisphere, retaining the original lat_lon column
- Set lat_value and lon_value columns to number type temporarily to support downstream math computation
- Computed a new latitude column: negated lat_value when lat_hemisphere equaled "S", otherwise kept lat_value as-is
- Computed a new longitude column: negated lon_value when lon_hemisphere equaled "W", otherwise kept lon_value as-is
- Deleted the intermediate columns lat_lon, lat_value, lat_hemisphere, lon_value, lon_hemisphere, leaving only the new latitude and longitude columns
- Joined the biosample_parsed table into the sra_run_metadata table using a full-outer join, bringing in the SRS and Title columns (first-match aggregation), and deleted the biosample_parsed table afterward
- Output as 1003894_v1_oysters_sra_run_metadata.csv

[[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

Barrett, P. D., Schneider, E. G., Grabowski, J. H., Hanley, T. C., McManus, M. C., Helt, W., Kinney, H., & Hughes, A. R. (2024). Evaluating Multiple Oyster Reef Restoration Practices Across Space and Time in Coastal Rhode Island. *Ecological Restoration*, 42(3), 193–204. <https://doi.org/10.3368/er.42.3.193>

Methods

Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A., & Cresko, W. A. (2013). Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22(11), 3124–3140. Portico. <https://doi.org/10.1111/mec.12354>

Software

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., & Durbin, R. (2011). The variant call format and VCFtools.

Bioinformatics, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>

Software

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.1207.3907> <https://doi.org/10.48550/arXiv.1207.3907>

Software

O’Leary, S. J., Puritz, J. B., Willis, S. C., Hollenbeck, C. M., & Portnoy, D. S. (2018). These aren’t the loci you’re looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology*, 27(16), 3193–3206. Portico. <https://doi.org/10.1111/mec.14792>

Methods

Parchman, T. L., Gompert, Z., Mudge, J., Schilkey, F. D., Benkman, C. W., & Buerkle, C. A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Molecular Ecology*, 21(12), 2991–3005. Portico. <https://doi.org/10.1111/j.1365-294x.2012.05513.x> <https://doi.org/10.1111/j.1365-294X.2012.05513.x>

Methods

Puritz, J. B., Guo, X., Hare, M., He, Y., Hillier, L. W., Jin, S., Liu, M., Lotterhos, K. E., Minx, P., Modak, T., Proestou, D., Rice, E. S., Tomlinson, C., Warren, W. C., Witkop, E., Zhao, H., & Gomez-Chiarri, M. (2024). A second unveiling: Haplotig masking of the eastern oyster genome improves population-level inference. *Molecular Ecology Resources*, 24(1). Portico. <https://doi.org/10.1111/1755-0998.13801>

Methods

Puritz, J. B., Hollenbeck, C. M., & Gold, J. R. (2014). *dDocent*: a RADseq, variant-calling pipeline designed for population genomics of non-model organisms. *PeerJ*, 2, e431. <https://doi.org/10.7717/peerj.431>

Truskey, S., Sotka, E., Grabowski, J., Kollars-Kjersten, N. M., Lotterhos, K. E., Schneider, E., & Hughes, A. R. (2025). Non-Random Mortality in an Experimental Oyster Restoration. *Evolutionary Applications*, 18(7). Portico. <https://doi.org/10.1111/eva.70128>

Results

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

Related Datasets

IsRelatedTo

Hughes, A. R., Truskey, S. (2026) **Genetic diversity metrics for oyster genetic clusters from an experimental oyster reef restoration in Ninigret Pond, Rhode Island (USA) in 2018-2020**. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-12 <http://lod.bco-dmo.org/id/dataset/1004745> [view at BCO-DMO]

Relationship Description: Related dataset with genetic diversity estimates at the oyster genetic cluster-level from restored reefs

Hughes, A. R., Truskey, S. (2026) **Genetic relatedness estimates calculated from SNP genotype data of oysters from an experimental oyster reef restoration in Ninigret Pond, Rhode Island (USA) in 2018-2020**. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-05 <http://lod.bco-dmo.org/id/dataset/1003777> [view at BCO-DMO]

Relationship Description: Related dataset with pairwise individual genetic relatedness estimates from restored reefs

Hughes, A. R., Truskey, S. (2026) **Oyster genetic assignment data from an experimental oyster reef restoration in Ninigret Pond, Rhode Island (USA) in 2017-2020**. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-07 <http://lod.bco-dmo.org/id/dataset/1004242> [view at BCO-DMO]

Relationship Description: Companion dataset with individual genetic cluster assignments for oysters

Hughes, A. R., Truskey, S. (2026) **Oyster morphometric, condition, and parasite infection data from an experimental oyster reef restoration in Ninigret Pond, Rhode Island (USA) in 2018-2020**. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-08-11 <http://lod.bco-dmo.org/id/dataset/1004478> [view at BCO-DMO]

Relationship Description: Morphometric, condition, and parasite infection rates of a sample subset

References

Northeastern University. Non-random mortality in an experimental oyster restoration. 2025/06. 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/PRJNA1280068>. NCBI:BioProject: PRJNA1280068.

Truskey, S. B., & Hughes, A. R. (2025). Non-random mortality in an experimental oyster restoration [Data set]. Northeastern University. <https://hdl.handle.net/2047/D20775066>

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

Parameters

--

Parameter	Description	Units
sample_name	Name of the sample	unitless
tissue	Type of tissue the sample was taken from	unitless
acc	SRA Run accession in the form of SRR##### (ERR or DRR for INSDC partners)	unitless
assay_type	Type of library (i.e. AMPLICON, RNA-Seq, WGS, etc)	unitless
batch	RADseq library batch in which an individual was processed and sequenced	unitless
bioproject	BioProject accession in the form of PRJNA##### (PRJEB##### or PRJDB##### for INSDC partners)	unitless
biosample	BioSample accession in the form of SAMN##### (SAMEA##### or SAMD##### for INSDC partners)	unitless
biosamplemodel	The BioSample package/model that was picked	unitless
SRS	NCBI SRA Sample accession (SRS)	unitless
datastore_filetype	Type of files available to download from SRA	unitless
datastore_provider	Locations of where the files are available to download from	unitless
datastore_region	Regions of where the data is located	unitless
experiment	The accession in the form of SRX##### (ERX or DRX for INSDC partners)	unitless
genotyped_individual	Simplified sample name that drops reef id from string	unitless
collection_date	The collection date of the sample	unitless
geo_loc_name_country	Name of the country where the sample was collected	unitless
geo_loc_name_country_continent	Name of the continent where the sample was collected	unitless

geo_loc_name	Full location of collection	unitless
latitude	Latitude where the sample was collected, North is positive	decimal degrees
longitude	Longitude where the sample was collected, West is negative	decimal degrees
isolate	Denotes the experimental block (local geographical source) from which an individual oyster sample was taken.	unitless
isolation_source	Denotes the reef ID within an experimental block (local geographical source) from which an individual oyster sample was taken.	unitless
instrument	Name of the sequencing instrument model	unitless
libraryselection	Library selection methodology (i.e. PCR, RANDOM, etc)	unitless
librarysource	Source of the biological data (i.e. GENOMIC, METAGENOMIC, etc)	unitless
organism	Scientific name of the organism that was sequenced (as found in the NCBI Taxonomy Browser)	unitless
platform	Name of the sequencing platform (i.e. ILLUMINA)	unitless
releasedate	The date on which the data was released	unitless
primary_genetic_assignment	Assigned genetic cluster for an individual using the primary genetic assignment set applied to all main text analyses (DAPC approach, SNP set filtered for MAF > 0.01, LD pruning, and Fall 2020 missing data). Label names correspond to the state of origin of the hatchery source associated with a given genetic cluster (gME, gMA, gNY, gVA).	unitless
excluded_from_reeflevel_analyses	Flag indicating whether an individual was excluded from reef-level analyses (i.e., if collected as initial source hatchery sample, extra reef samples from 2019 or 2021, or from reefs shown to have started as single source reefs in 2018). NA if sample not excluded.	unitless
Title	Biosample title	unitless

Instruments

Dataset-specific Instrument Name	Illumina HiSeq 2500
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	The initial and fall 2020 batches were sequenced on a single lane of the Illumina HiSeq 2500 at Tufts University Core Facility Genomics; the fall 2018 batch was sequenced on two lanes of the Illumina NovaSeq 6000 at the University of Texas at Austin Genomic Sequencing and Analysis Facility.
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 6000
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	The initial and fall 2020 batches were sequenced on a single lane of the Illumina HiSeq 2500 at Tufts University Core Facility Genomics; the fall 2018 batch was sequenced on two lanes of the Illumina NovaSeq 6000 at the University of Texas at Austin Genomic Sequencing and Analysis Facility.
Generic Instrument Description	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

Dataset-specific Instrument Name	
Generic Instrument Name	Manual Biota Sampler
Dataset-specific Description	In fall 2018, we haphazardly sampled live oysters from each reef on scuba or snorkel (N = 512 individuals total, 32 per reef). Oysters were put on ice and transported to the Northeastern University Marine Science Center where they were held at -80°C until DNA extraction.
Generic Instrument Description	"Manual Biota Sampler" indicates that a sample was collected in situ by a person, possibly using a hand-held collection device such as a jar, a net, or their hands. This term could also refer to a simple tool like a hammer, saw, or other hand-held tool.

Dataset-specific Instrument Name	scuba
Generic Instrument Name	Self-Contained Underwater Breathing Apparatus
Dataset-specific Description	In fall 2018, we haphazardly sampled live oysters from each reef on scuba or snorkel (N = 512 individuals total, 32 per reef). Oysters were put on ice and transported to the Northeastern University Marine Science Center where they were held at -80°C until DNA extraction.
Generic Instrument Description	The self-contained underwater breathing apparatus or scuba diving system is the result of technological developments and innovations that began almost 300 years ago. Scuba diving is the most extensively used system for breathing underwater by recreational divers throughout the world and in various forms is also widely used to perform underwater work for military, scientific, and commercial purposes. Reference: https://oceanexplorer.noaa.gov/technology/technical/technical.html

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

Project Information

CAREER: Linking genetic diversity, population density, and disease prevalence in seagrass and oyster ecosystems (Seagrass and Oyster Ecosystems)

Coverage: Coastal New England

NSF Award Abstract:

Disease outbreaks in the ocean are increasing, causing losses of ecologically important marine species, but the factors contributing to these outbreaks are not well understood. This 5-year CAREER project will study disease prevalence and intensity in two marine foundation species - the seagrass *Zostera marina* and the Eastern oyster *Crassostrea virginica*. More specifically, host-disease relationships will be explored to understand how genetic diversity and population density of the host species impacts disease transmission and risk. This work will pair large-scale experimental restorations and smaller-scale field experiments to examine disease-host relationships across multiple spatial scales. Comparisons of patterns and mechanisms across the two coastal systems will provide an important first step towards identifying generalities in the diversity-density-disease relationship. To enhance the broader impacts and utility of this work, the experiments will be conducted in collaboration with restoration practitioners and guided by knowledge ascertained from key stakeholder groups. The project will support the development of an early career female researcher and multiple graduate and undergraduate students. Students will be trained in state-of-the-art molecular techniques to quantify oyster and seagrass parasites. Key findings from the surveys and experimental work will be incorporated into undergraduate courses focused on Conservation Biology, Marine Biology, and Disease Ecology. Finally, students in these courses will help develop social-ecological surveys and mutual learning games to stimulate knowledge transfer with stakeholders through a series of workshops.

The relationship between host genetic diversity and disease dynamics is complex. In some cases, known as a dilution effect, diversity reduces disease transmission and risk. However, the opposite relationship, known as the amplification effect, can also occur when diversity increases the risk of infection. Even if diversity directly reduces disease risk, simultaneous positive effects of diversity on host density could lead to amplification by increasing disease transmission between infected and uninfected individuals. Large-scale field restorations of seagrasses (*Zostera marina*) and oysters (*Crassostrea virginica*) will be utilized to test the effects of host genetic diversity on host population density and disease prevalence/intensity. Additional field experiments independently manipulating host genetic diversity and density will examine the mechanisms leading to dilution or amplification. Conducting similar manipulations in two marine foundation species - one a clonal plant and the other a non-clonal animal - will help identify commonalities in the diversity-density-disease relationship. Further, collaborations among project scientists, students, and stakeholders will enhance interdisciplinary training and help facilitate the exchange of information to improve management and restoration efforts. As part of these efforts, targeted surveys will be used to document the perceptions and attitudes of managers and restoration

practitioners regarding genetic diversity and its role in ecological resilience and restoration.

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

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-1652320

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