

Parent eastern oysters (*Crassostrea virginica*) from wild and hatchery populations along U.S. Atlantic and Gulf coasts (collected 2022), used as broodstock in a field experiment (2023-2025)

Website: <https://www.bco-dmo.org/dataset/1008652>

Data Type: Other Field Results, experimental

Version: 1

Version Date: 2026-09-30

Project

» [CAREER: Evaluation of machine learning algorithms for understanding and predicting adaptation to multivariate environments with a Model Validation Program \(MVP\)](#) (Model Validation Program)

Contributors	Affiliation	Role
Lotterhos, Katie	Northeastern University	Principal Investigator
Armstrong, Madison	Northeastern University	Scientist
Carnegie, Ryan	Virginia Institute of Marine Science (VIMS)	Scientist
Eppley, Madeline	Northeastern University	Scientist
Rumberger, Camille	Northeastern University	Scientist
Small, Jessica	The College of William & Mary	Scientist
Bajaj, Kiran	Northeastern University	Student
K. Eaton, Anna	Northeastern University	Student
Mongillo, Nicole	Northeastern University	Student
Segnitz, Zea	Northeastern University	Student
York, Amber D.	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract

This project is related to the CviMVP (Model Validation Program) project, which aims to assess Machine Learning Algorithms (MLAs) for understanding and predicting adaptation of organisms to multivariate environments from their DNA sequences. The data here is related to a field experiment validating the results of MLAs predicting responses to common garden conditions at two sites in the Chesapeake Bay, Lewisetta and York River. The metadata includes information on source population, phenotype data and experimental common garden. Briefly, the experiment involved spawning oysters from two selection lines and six wild populations, and creating spawning treatments including monocultures of all these groups as well as two polyculture treatments. Juvenile oysters were then deployed at two common gardens at Lewisetta, VA and York River, VA and monitored over the course of two years. This dataset was used to assign parents to offspring to determine parental success in polyculture treatments.

Table of Contents

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

Coverage

Location: U.S. Atlantic and Gulf coasts, including Chesapeake Bay

Spatial Extent: N:44.02 E:-69.53 S:28.1 W:-97.17

Temporal Extent: 2022-09-01 - 2025-05-01

Dataset Description

Organism identifiers:

Eastern oyster, *Crassostrea virginica*, urn:lsid:marinespecies.org:taxname:140657

Methods & Sampling

Experimental Treatments

The monoculture treatments included six wild populations and two selectively bred populations (DEBY and LOLA) owned by the Aquaculture Genetics and Breeding Technology Center (ABC, Virginia Institute of Marine Science). Wild populations consisted of Maine, New Hampshire, Virginia, Florida, Louisiana, and Texas oysters. Sites will be referred to as acronyms in the remainder of this paper: ME, NH, VA, FL, LA and TX, respectively. Such locations span the native range of the eastern oyster, which is structured by two distinct genetic clusters ("Atlantic" and "Gulf") separated by the Florida Peninsula (Puritz et al., 2022; Reeb and Avise, 1990). The LOLA selection line is considered local to Lewisetta and DEBY is considered local to the York River, for this study. Oysters for each monoculture were sourced September–November 2022 and conditioned at the Acuff Center for Aquaculture at VIMS until spawning in May 2023.

For each treatment, oysters were strip spawned and sperm from ten males per population were individually crossed with a pool of ten eggs from equally represented females to limit sperm competition (Levitan, 2004). Sperm from each male was then crossed with eggs from all females, producing new combinations and increasing genetic diversity in a polyculture treatment, "LARMIX". The larvae were raised following standard hatchery (days 0–30) and nursery (days 30–77) methods. On day 77, a second polyculture treatment, "SEEDMIX", was created by equally mixing seed from each monoculture to increase genetic diversity while maintaining original genotypes.

Field Deployment

Seed from each treatment were split into six bags (approximately 1,200 per bag). Three replicate bags were deployed in bottom cages at east common garden site in August 2023 and assessed in November 2023 (time point 1), May 2024 (time point 2), November 2024 (time point 3), and May 2025 (final time point). At the first assessment, 40 oysters per bag were individually tagged. At each time point, tagged oysters were determined to be alive or dead. Note that not all tags were found at every assessment, some were missed due to debris on the oyster shell and others had become separated from the oyster in between assessments. For tagged oysters that could be identified, shell length (longest distance from the hinge to the bill) and width (longest distance perpendicular to the length) were measured. To reduce density-dependent effects, bags were thinned to 225 oysters at the second assessment (Marshall and Dunham, 2013).

Genetic data

Gill tissue was taken from each parent ($n = 160$, 83 dams and 77 sires) after spawning. At the second assessment, 267 tagged juvenile oysters were chosen at random for a nonlethal biopsy to collect gill tissue. All samples were stored in 95% ethanol at -80°C . DNA was extracted using the Qiagen DNeasy Blood and Tissue Kit and shipped to Neogen Genomics in Lincoln, NE for genotyping on a 200K ThermoFisher Affymetrix Axiom SNP array derived from a 600K array assigned to the haplotig-masked *Crassostrea virginica* reference genome (Gómez-Chiarri et al., 2015; Guo et al., 2023; Modak et al., 2021; Puritz et al., 2024). The sequencing pipeline is outlined in Bajaj and Mongillo et al. (2026). Raw genotypes were processed and filtered in R v4.4.2 and a full, unimputed, SNP set was used for further analysis.

Data Processing Description

NA

BCO-DMO Curation Notes

Curation Actions Performed on Data

- Loaded parents.csv into table named 1008652_v1_cvimvp-parents, with header row 1, and "" and "NA" set as missing value identifiers.
- Deleted column Mortality_Date (no values in this column and it was not included in the column descriptions by the data submitter).
- Renamed columns Pass.Fail to Pass_Fail, QC.call_rate to QC_call_rate, Average.call.rate.for.passing.samples to Average_call_rate_for_passing_samples, Sample.FileName to Sample_Filename, and QC.het_rate to QC_het_rate, replacing periods with underscores to conform field names to naming convention
- Set data types for all columns: Alive as boolean; Average_call_rate_for_passing_samples, DQC, QC_call_rate, QC_het_rate, Shell_length_cm, Shell_width_cm, call_rate, and het_rate as number; Sperm_Motility as integer; remaining columns (Buffer, Extraction_plan_Key, Extraction_plan_Label, ID_SiteDate, Key_Vial, Key_box, Key_ind, Label_ind, New_Growth, Pass_Fail, PlateID_Key, Sample_Filename, Sex, SiteDate_Key, Tissue, Vial_Label, clean_ID, pops_Key) as string
- Updated column metadata with descriptions, standard parameter name IDs, and supplied units for all columns, including notes describing Alive, Average_call_rate_for_passing_samples, Buffer, DQC, Extraction_plan_Key, Extraction_plan_Label, ID_SiteDate, Key_Vial, Key_box, Key_ind, Label_ind, New_Growth, Pass_Fail, PlateID_Key, QC_call_rate, QC_het_rate, Sample_Filename, Sex, Shell_length_cm, Shell_width_cm, SiteDate_Key, Sperm_Motility, Tissue, Vial_Label, call_rate, clean_ID, and het_rate
- Updated package-level metadata with per-column summary statistics (counts, minimum, and maximum values) for all columns
- Saved table as 1008652_v1_cvimvp-parents.csv

Curation Actions Performed on Metadata

- * Standard metadata entry and formatting of provided metadata.
- * Data manager set the temporal and spatial bounds to the same lat lon and time range was used as the related Environmental dataset (See related datasets) which included the broodstock source population collection time and locations.

Known Issues [Potentially Impacting Reuse]

- * Information for sites (lat,lons, etc) and information for source populations of parental broodstock genotyped were not included in this dataset directly. See Related Datasets where that information was provided. However, the exact way to associate that data with this dataset was not explicitly stated (shared keys not explained). However there is an implied association between site IDs and source populations/spawning treatment IDs due to the shared use of state abbreviations within the values (e.g. source population identified by "FL-KingPlan" may be from the FL site "W3_FL, 30.44003,-81.43638").

Problem Description

NA

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

Related Publications

Bajaj, K. E., Mongillo, N., Eppley, M. G., Rumberger, C. A., Segnitz, Z., Katsuki, S., Carnegie, R., Small, J., & Lotterhos, K. E. (2026). Contrasting effects of geographic distance, environmental distance, and intraspecific diversity on the performance of a marine invertebrate in common gardens.

<https://doi.org/10.64898/2026.04.02.716183>

Results

Guo, X., Puritz, J. B., Wang, Z., Proestou, D., Allen, S., Small, J., Verbyla, K., Zhao, H., Haggard, J., Chriss, N., Zeng, D., Lundgren, K., Allam, B., Bushek, D., Gomez-Chiarri, M., Hare, M., Hollenbeck, C., La Peyre, J., Liu, M., et al. (2023). Development and Evaluation of High-Density SNP Arrays for the Eastern Oyster *Crassostrea virginica*. *Marine Biotechnology*, 25(1), 174–191. <https://doi.org/10.1007/s10126-022-10191-3>

Methods

Gómez-Chiarri, M., Warren, W. C., Guo, X., & Proestou, D. (2015). Developing tools for the study of molluscan immunity: The sequencing of the genome of the eastern oyster, *Crassostrea virginica*. *Fish & Shellfish Immunology*, 46(1), 2-4. <https://doi.org/10.1016/j.fsi.2015.05.004>

Methods

Levitan, Don R. (2004). Density-Dependent Sexual Selection in External Fertilizers: Variances in Male and Female Fertilization Success along the Continuum from Sperm Limitation to Sexual Conflict in the Sea Urchin *Strongylocentrotus franciscanus*. *The American Naturalist*, 164(3), 298-309. <https://doi.org/10.1086/423150>

Methods

Marshall, R. D., & Dunham, A. (2013). Effects of culture media and stocking density on biofouling, shell shape, growth, and survival of the Pacific oyster (*Crassostrea gigas*) and the Manila clam (*Venerupis philippinarum*) in suspended culture. *Aquaculture*, 406-407, 68-78. <https://doi.org/10.1016/j.aquaculture.2013.05.003>

Methods

Modak, T. H., Literman, R., Puritz, J. B., Johnson, K. M., Roberts, E. M., Proestou, D., Guo, X., Gomez-Chiarri, M., & Schwartz, R. S. (2021). Extensive genome-wide duplications in the eastern oyster (*Crassostrea virginica*). *Philosophical Transactions of the Royal Society B*, 376(1825). <https://doi.org/10.1098/rstb.2020.0164>

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., Zhao, H., Guo, X., Hare, M. P., He, Y., LaPeyre, J., Lotterhos, K. E., Lundgren, K. M., Modak, T., Proestou, D., Rawson, P., Fernandez Robledo, J. A., Weedop, K. B., Witkop, E., & Gomez-Chiarri, M. (2022). Nucleotide and structural polymorphisms of the eastern oyster genome paint a mosaic of divergence, selection, and human impacts. <https://doi.org/10.1101/2022.08.29.505629>

Methods

Reeb, C. A., & Avise, J. C. (1990). A genetic discontinuity in a continuously distributed species: mitochondrial DNA in the American oyster, *Crassostrea virginica*. *Genetics*, 124(2), 397-406. <https://doi.org/10.1093/genetics/124.2.397>

Methods

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

Related Datasets

IsRelatedTo

Bajaj, K., Eppley, M., Rumberger, C., Mongillo, N., Lotterhos, K., Small, J., Carnegie, R., Katsuki, S., Segnitz, Z. (2026) **SNP genotypes of Eastern oysters (*Crassostrea virginica*) from wild and hatchery populations along U.S. Atlantic and Gulf coasts (collected 2022), used as broodstock in a field experiment (2023-2025)**. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-10-05 <http://lod.bco-dmo.org/id/dataset/998738> [\[view at BCO-DMO\]](#)
Relationship Description: The experimental results dataset (1007522) includes fitness data and metadata from a common garden experiment run as part of the CviMVP project from 2023-2025. In the experiment juvenile oysters were transplanted to two sites to assess fitness across environments. These juveniles were spawned from a set of parental oysters, and data and metadata on that parental generation (used to produce the juveniles for this experiment) are contained in a separate dataset (1008652). Additional genotype information for the parent oysters is provided in dataset (998738). Environmental data from the experiment is in dataset 1007572.

Rumberger, C., Lotterhos, K., Eppley, M., Small, J., Katsuki, S., Mongillo, N., Bajaj, K., Carnegie, R., Segnitz, Z. (2026) **Eastern oyster (*Crassostrea virginica*) mortality, growth, and tissue sample metadata from the common garden experiments run as part of CviMVP project from 2023-2025 in the Chesapeake Bay**. Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-10-06 <http://lod.bco-dmo.org/id/dataset/1007522> [\[view at BCO-DMO\]](#)
Relationship Description: The experimental results dataset (1007522) includes fitness data and metadata from a common garden experiment run as part of the CviMVP project from 2023-2025. In the experiment juvenile

oysters were transplanted to two sites to assess fitness across environments. These juveniles were spawned from a set of parental oysters, and data and metadata on that parental generation (used to produce the juveniles for this experiment) are contained in a separate dataset (1008652). Additional genotype information for the parent oysters is provided in dataset (998738). Environmental data from the experiment is in dataset 1007572.

Rumberger, C., Lotterhos, K., Small, J., Eppley, M., Carnegie, R., Mongillo, N., Bajaj, K., Segnitz, Z. (2026) **Environmental data from sites (source populations, and experimental sites) from common garden experiments with Eastern oyster (*Crassostrea virginica*) run as part of CviMVP project from 2023-2025 in the Chesapeake Bay.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2026-10-06 <http://lod.bco-dmo.org/id/dataset/1007572> [[view at BCO-DMO](#)]

Relationship Description: The experimental results dataset (1007522) includes fitness data and metadata from a common garden experiment run as part of the CviMVP project from 2023-2025. In the experiment juvenile oysters were transplanted to two sites to assess fitness across environments. These juveniles were spawned from a set of parental oysters, and data and metadata on that parental generation (used to produce the juveniles for this experiment) are contained in a separate dataset (1008652). Additional genotype information for the parent oysters is provided in dataset (998738). Environmental data from the experiment is in dataset 1007572.

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

Parameters

Parameter	Description	Units
Label_ind	4 digit code for each individual oyster.	unitless
Key_ind	Unique identifier key corresponding to each individual.	unitless
clean_ID	Label corresponding to each individual with plate well_project label_plate_ number_individual label_individual key.	unitless
Vial_Label	Label corresponding to each sample with project label_number identifier.	unitless
Sample_Filename	Label corresponding to each individual with plate well_project label_plate_ number_individual label_individual key with additional information on plate number to match vcfile.	unitless
Pass_Fail	Pass or Fail for the individual QC from the sequencing facility.	unitless
DQC	DQC, or Dish QC, is a pre-genotyping quality control metric used to assess the quality of a hybridized Affymetrix SNP array before full genotype calling is performed.	unitless

QC_call_rate	QC call rate is a sample-level quality control metric that measures the proportion of SNPs on the array that were successfully genotyped (i.e., assigned a genotype call) for a given sample on the Affymetrix SNP Array. A consistent benchmark across labs and studies and used to determine pass/fail.	unitless
call_rate	This also measures the proportion of SNPs on the array that were successfully genotyped Affymetrix SNP Array, but is specific to the study.	unitless
QC_het_rate	Heterozygosity rate measures the proportion of SNPs for a given sample that are called as heterozygous (AB) rather than homozygous (AA or BB) on the Affymetrix SNP Array.	unitless
het_rate	Heterozygosity rate measures the proportion of SNPs for a given sample that are called as heterozygous (AB) rather than homozygous (AA or BB) on the Affymetrix SNP Array. This is specific to the study and used for later filtering steps.	unitless
Average_call_rate_for_passing_samples	Average.call.rate.for.passing.samples is a batch-level (plate/run-level) QC metric rather than an individual sample metric. It summarizes the overall genotyping performance across all samples in a batch that passed individual QC filters.	unitless
Key_box	Unique identifier key corresponding to each box the tissue samples were located in.	unitless
Key_Vial	Unique identifier key corresponding to each vial of tissue sample.	unitless
Tissue	Type of tissue taken for final tissue sample (gill or mantle).	unitless
Buffer	Preservative in the vial the final tissue sample was placed in.	unitless
Extraction_plan_Key	Unqiue identifier for extraction sample.	unitless
PlateID_Key	Unique identifier key corresponding to each plate of DNA.	unitless
ID_SiteDate	Source population label.	unitless

SiteDate_Key	Unique identifier key corresponding to each source population.	NA
Extraction_plan_Label	Label corresponding to each extraction sample with plate well_project label_plate_ number_individual label_individual key.	unitless
Sex	Sex of the oyster based on visual inspection of gametes for ripe oysters. If the eggs were visible in a clear beaker held up to the light, we labeled it as a female. If the beaker was cloudy we labeled it as a male.	unitless
Shell_length_cm	Length of each oyster (the longest distance from the hinge to the bill).	centimeters (cm)
Shell_width_cm	Width of each oyster (the longest distance perpendicular to the length measurement).	centimeters (cm)
pops_Key	Unique identifier key corresponding to each source population.	unitless
Alive	Information on if the individual oyster was alive or dead (binary, TRUE or FALSE).	unitless
New_Growth	Information if shells had evidence of new growth for assessment of health (binary assignment, Yes or No).	unitless
Sperm_Motility	Information on how mobile the sperm of each individual was (range with 4 is fastest, 1 is slowest).	unitless

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

Project Information

CAREER: Evaluation of machine learning algorithms for understanding and predicting adaptation to multivariate environments with a Model Validation Program (MVP) (Model Validation Program)

Coverage: East coast of North America

NSF Award Abstract:

Environmental change can be rapid and involve multiple aspects of the environment changing at the same time, such as warming and increased disease pressure. Rapid environmental change threatens the productivity of aquaculture and crops on which humans depend. Predicting organisms' vulnerabilities to rapid and multifactor environmental change, however, is a major scientific challenge. A hurdle to addressing this challenge arises from the complex and non-intuitive ways that organisms adapt, through changes at the level of the DNA sequence, to many environmental stresses at the same time. Thus, there is a need for new approaches to understand and predict adaptation in multivariate environments. To address this need, this project integrates research and education with a Model Validation Program (MVP). The research is developing and evaluating

Machine Learning Algorithms (MLAs) for understanding and predicting adaptation of organisms to multivariate environments from their DNA sequences. To evaluate MLAs, this research combines both data simulation and an empirical test in the field with the Eastern Oyster, which provide important ecosystem services and support a multi-million dollar industry. For oysters, this research is studying how temperature, disease pressure, and salinity interact with evolutionary history to determine fitness in the field. This research advances efforts toward addressing the major scientific challenge of predicting adaptation in complex environments by integrating concepts across the frontiers of marine, evolutionary, and statistical sciences in a new way. Machine learning and model validation are not traditionally taught in the marine and environmental sciences, but are becoming increasingly relevant to these fields. As part of a broader education program, this research is developing MVP Learning Modules for high school students and undergraduates, which help students build the foundational knowledge they need to critically evaluate and apply models. Modules are being disseminated to hundreds of students in the greater Boston area and are being made available online for widespread use. The MVP mentoring program is training graduate students, undergraduates, and high school students in marine evolutionary ecology, statistical genomics, and machine learning. This research addresses a pressing societal need to more informatively match genotypes to environments for restoration, farming, and assisted gene flow efforts. Results are being disseminated to stakeholders in the oyster industry.

The goal of this research is to evaluate if MLAs, which can model non-linearities, can be used to understand and predict adaptation to multivariate environments under a wide range of scenarios. In Objective 1, the Principal Investigator (PI) is creating simulated datasets with different aspects of realism, and using them to evaluate and refine the MLAs. This novel set of simulations is studying genome evolution under high gene flow in complex, multivariate environments. In Objective 2, the PI is building on their expertise with the Eastern oyster to evaluate the MLAs in a field setting. The PI is first developing a comprehensive seascape genomic dataset and using it to train MLAs to predict an individual's multivariate environment based on a single nucleotide polymorphism genotype. Then, the PI is testing if the MLA prediction can predict the fitness of different genotypes from across the species range when raised in common garden field conditions. In Objective 3, the PI is integrating research and education by using the data obtained from Objs. 1 and 2 to develop a series of original "MVP Learning Modules" with interactive web apps for persons at different levels of understanding, using the relatable example of an oyster restoration project. This research lays the foundation for future studies by producing datasets that could become classical examples for developing and benchmarking innovative modeling approaches.

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-2043905

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