Eastern oyster parent genotypes

Website: https://www.bco-dmo.org/dataset/998738
Data Type: experimental
Version: 1
Version Date: 2026-05-18

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
Small, Jessica	Virginia Institute of Marine Science (VIMS)	Co-Principal Investigator
Carnegie, Ryan	Virginia Institute of Marine Science (VIMS)	Scientist
Bajaj, Kiran	Northeastern University	Student
Eppley, Madeline	Northeastern University	Student
Katsuki, Shelley	Virginia Institute of Marine Science (VIMS)	Student
Mongillo, Nicole	Northeastern University	Student
Rumberger, Camille	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. 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 (mixed) 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. In-depth mortality and phenotype data were collected every six months and used to assess the fitness of individuals from different spawn treatments across the two gardens. This dataset includes broodstock (parent) genotype data generated on a 200K ThermoFisher Affymetrix Axiom SNP array derived from a 600K array (Gómez-Chiarri et al., 2015; Guo et al., 2023; Modak et al., 2021; Puritz et al., 2024) aligned to the haplotig-masked Crassostrea virginica reference genome (C_virginica-3.0, GCA_002022765.4). The genotype data are archived as two files: (1) a SNP genotype matrix containing processed, imputed genotype calls per individual encoded in 0/1/2 dosage format (homozygous reference, heterozygous, homozygous alternate), and (2) a SNP metadata file providing per-locus information including SNP identifiers, genomic positions, gene annotations, Gene Ontology terms, and population genetic statistics from OutFLANK outlier detection (FST, expected heterozygosity, p-values, q-values, and outlier flags).

Table of Contents

Coverage

Location: Atlantic and Gulf coasts of the United States
Spatial Extent: **N**:43.986 **E**:-69.55 **S**:28.084 **W**:-97.201
Temporal Extent: 2023-05-01 - 2025-05-01

Dataset Description

Acronyms:

SNP = Single-nucleotide Polymorphism
MLA = Machine Learning Algorithm
LD = Linkage Disequilibrium
NCBI = National Center for Biotechnology Information
DNA = Deoxyribonucleic acid
MVP = Model Validation Program
ABC = Aquaculture Genetics and Breeding Technology Center
VIMS = Virginia Institute of Marine Science
GO = Gene Ontology

Methods & Sampling

Source populations for monoculture treatments included six wild populations and two proprietary broodstock lines (DEBY and LOLA) from the Aquaculture Genetics and Breeding Technology Center (ABC), Virginia Institute of Marine Science (VIMS). Wild populations spanned the species' native range, which is structured into distinct genetic clusters separated by the Florida peninsula (Reeb & Avise, 1990, Puritz et al., 2022). Populations included oysters from: a variable salinity site in Texas (W1-TX), a low salinity site in Louisiana (W2-LA), a high salinity site on the east coast of Florida (W3-FL), a moderate salinity site in the James River, Virginia (W4-VA), a variable salinity site in New Hampshire (W5-NH), and a high salinity site in Maine (W6-ME). W4-VA was considered local to both common garden sites due to its geographic proximity and intermediate salinity.

Gill tissue was collected from each parent (n = 160) after spawning and stored in 95% ethanol at -80°C. DNA was extracted using the Qiagen DNeasy Blood and Tissue Kit and shipped to Neogen Genomics (Lincoln, NE) for genotyping on a 200K ThermoFisher Affymetrix Axiom SNP array derived from a 600K array (Gómez-Chiarri et al., 2015; Guo et al., 2023; Modak et al., 2021; Puritz et al., 2024) aligned to the haplotig-masked *Crassostrea virginica* reference genome (C_virginica-3.0, GCA_002022765.4).

Raw genotypes were processed and filtered in R v4.4.2 (R Core Team 2025). Missing SNPs were imputed using LEA v3.6.0 (Frichot & François, 2015) with K = 2 ancestral groups to produce a full SNP set. We thinned SNPs for linkage disequilibrium (LD) for population structure analysis and neutral parameterization of genome scans (Lotterhos, 2019) using bigsnpr v1.11.3 (Privé et al., 2018). The archived genotype matrix contains the processed, imputed genotype calls per individual encoded in 0/1/2 integer format (homozygous reference, heterozygous, and homozygous alternate, respectively). The accompanying SNP metadata file provides per-locus descriptions including genomic position, gene identifiers, functional descriptions, Gene Ontology terms, and population genetic statistics derived from OutFLANK (FST, expected heterozygosity, p-values, q-values, and outlier flags), as well as a flag indicating inclusion in the LD-thinned dataset, enabling straightforward filtering for only LD-thinned SNPs.

Data Processing Description

Raw genotypes were processed and filtered in R v4.4.2 (R Core Team 2025; Supp. Methods: genotype filtering). Missing SNPs were imputed using LEA v3.6.0 (Frichot & François, 2015) with K = 2 ancestral groups to produce a *full SNP set*. We thinned SNPs for linkage disequilibrium (LD) for population structure analysis and neutral parameterization of genome scans (Lotterhos, 2019) using bigsnpr v1.11.3 (Privé et al., 2018).

https://github.com/DrK-Lo/MVP-H2F-HatcheryField

BCO-DMO Processing Description

- Loaded data from "Exp_parents_SNP_metadata.csv" into table "998738_v1_oyster-genotype-annotations" using CSV format with row 1 as header; treated empty strings and "nd" as missing values
- Renamed column "AX-ID" to "AX_ID"
- Set data types for all 29 columns: AX_ID, Affx_ID, Group, SNPType, Sequence, cust_id, gene_description, gene_id, go_ids, go_terms, mutID, on_oystercv, organism, scaffold, OutlierFlag, thinned_dataset as string; Chromosome, Position, Rank, Replicates, Tile_Std, Tile_max, Tile_v3 as integer; FST, He, chrom_position, pvalues, pvaluesRightTail, qvalues as number
- Output final table as "998738_v1_oyster-genotype-annotations.csv"

- Provided matrix file Exp_parents_full_SNP_matrix.rds attached as a data file directly. [not imported as a table]

Problem Description

NA

[table of contents | back to top]

Related Datasets

Software

(n.d.). MVP-H2F-HatcheryField: MVP experimental data hatchery to field [Software repository]. GitHub. https://github.com/DrK-Lo/MVP-H2F-HatcheryField

[table of contents | back to top]

Parameters

Parameter	Description	Units
Affx_ID	Thermofisher SNP identifiers that refer to the specific probe or probe set as it was designed for the array	unitless
AX_ID	Thermofisher analysis-ready SNP identifiers that are assigned after the array is finalized	unitless
Chromosome	Chromosome where SNP occurs on assembly GCF_002022765.2_C_virginica-3.0_genomic	unitless
Position	Chromosomal position where SNP occurs on assembly GCF_002022765.2_C_virginica-3.0_genomic	unitless
FST	Fixation index for the locus calculated by OutFLANK (Whitlock and Lotterhos 2015)	unitless

He	Expected heterozygosity for the locus calculated by OutFLANK (Whitlock and Lotterhos 2015)	unitless
pvalues	Two-tailed p-value for the test of neutrality for the locus calculated by OutFLANK (Whitlock and Lotterhos 2015)	unitless
pvaluesRightTail	Right-tailed p-value for the test of neutrality for the locus calculated by OutFLANK (Whitlock and Lotterhos 2015)	unitless
qvalues	q-value for the test of neutrality for the locus based on the right-tailed p-value calculated by OutFLANK (Whitlock and Lotterhos 2015)	unitless
OutlierFlag	True (T) or False (F) if the SNP was detected as an outlier by OutFLANK (Whitlock and Lotterhos 2015)	unitless
gene_id	Generic locus IDs (NCBI) for gene(s) where SNP occurs	unitless
gene_description	Functional description(s) of gene(s) where SNP occurs	unitless
go_ids	Unique seven-digit identifier for a gene ontology term	unitless
go_terms	Description for a gene ontology term	unitless
SNPType	K - G/T M - A/C R - A/G (most common) S - C/G (rarest) W - A/T Y - C/T (most common)	unitless
Tile_Std	Affymetrix parameter for SNP chip design	unitless
Tile_max	Affymetrix parameter for SNP chip design	unitless
Tile_v3	SNPs tiled on the 200K chip	unitless
Replicates	Affymetrix parameter for SNP chip design	unitless
Rank	Affymetrix parameter for SNP chip design	unitless
Group	pathogen detection CN probesets, content of Axiom_OysterCV, additional well-performing markers from screen	unitless
organism	Genus species	unitless

cust_id	Annotation	unitless
on_oystercv	on Axiom_OysterCV developed be Breeding Consortium	unitless
mutID	Mutation ID Affymetrix parameter for SNP chip design	unitless
chrom_position	Chromosome location and position of SNP on the 200K array; separated by a '.'	unitless
scaffold	Scaffold on assembly GCF_002022765.2_C_virginica-3.0_genomic	unitless
Sequence	71mer sequence	unitless
thinned_dataset	True or False if the SNP occurs in the linkage-disequilibrium thinned SNP dataset	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]