Output model data from paradox of adaptive trait clines with non-clinal patterns in the underlying genes (Model Validation Program project)

Website: https://www.bco-dmo.org/dataset/889769
Data Type: model results
Version: 1
Version Date: 2023-02-13

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, Contact
Newman, Sawyer	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
Background: Multivariate climate change presents an urgent need to understand how species adapt to complex environments. Population genetic theory predicts that loci under selection will form monotonic allele frequency clines with their selective environment, which has led to the wide use of genotype-environment associations (GEAs). This study used a novel set of In silico simulations to elucidate the conditions under which allele frequency clines are more or less likely to evolve as multiple quantitative traits adapt to multivariate environments. Zenodo archive of GitHub Repository of all code used to create the simulations. Every directory includes a README describing the code, and metadata files are included for the archived outputs. Modeling code details: Code was developed 2020-2022 Simulation code was developed in SLiM, recapitated in pyslim, filtered with vcftools, and analyzed with R. Code was developed by K. E. Lotterhos (PI)

Table of Contents

Coverage

Temporal Extent: 2020 - 2022

Dataset Description

See metadata files associated with simulation outputs in the repository.

The associated modeling code is archived at 10.5281/zenodo.7622893.

The tutorial associated with the publication is published at https://marineomics.github.io/RDAtraitPredictionTutorial.html.

Methods & Sampling

Landscapes and demographies

All simulations consisted of 100 demes arranged on a 10 x 10 landscape grid. The 15 levels of landscape-demography were broadly divided into three landscape categories: (i) a stepping stone landscape with latitudinal and longitudinal selective clines (Stepping-Stone Clines, the most commonly simulated scenario in testing methods) (Coop et al. 2010; Frichot et al. 2013; Günther & Coop 2013; de Villemereuil et al. 2014; Lotterhos & Whitlock 2015; Rellstab et al. 2015; Gautier 2015; Forester et al. 2016, 2018), (ii) a stepping stone landscape with one latitudinal cline and one non-linear longitudinal mountain range (Stepping-Stone Mountain, which left the potential for unique architectures to arise to the same selective pressure at different geographic locations), and (iii) an estuary landscape with a latitudinal and longitudinal selective clines (Estuary Clines, which simulated repeated independent bouts of adaptation analogous to oysters or sticklebacks that repeatedly colonize and adapt to isolated freshwater environments connected by gene flow in the marine environment). For simplicity, I refer to the latitudinal environment as Temperature and the longitudinal environment as Env2.

In summary, Stepping-Stone Mountain had a different environmental pattern than Stepping-Stone Clines but the same demography, while Estuary Clines had the same environmental pattern as Stepping-Stone Clines but different demography. The demographic parameters were chosen such that different landscapes achieved similar levels of neutral genetic differentiation and local adaptation. Within each of the three landscapes, 5 demographies were simulated that described the migration rates and effective population sizes on the landscape (see Supplemental Methods).

See the Supplemental Methods for a description of the multivariate continuous space simulations with 6 traits.

Genetic map

The Wright-Fisher simulations were based on a previously published quantitative genetic model and a genetic map (Lotterhos 2019). The genome consisted of 20 linkage groups each with 50,000 sites. The scaled recombination rate (N metapop r = 0.01) gave a resolution of 0.001 cM between proximate bases and a total length of 50 cM for each linkage group. This resolution mimicked a SNP chip, in which SNPs were collected across a large genetic map (Lotterhos 2019). The population-scaled neutral mutation rate was (N metapop μ= 0.001). QTNs could evolve on the first 10 linkage groups, while on the second 10 linkage groups only neutral loci could evolve.

Genetic Architecture and Stabilizing Selection

Mutation: Quantitative trait nucleotides (QTNs) contributed additively to the optimal phenotype for each individual without dominance. Three genic levels were simulated: oligogenic (few loci of large effect on the trait), moderately polygenic (dozens to hundreds of loci with intermediate effects), and highly polygenic (hundreds of loci with small effects). For QTN mutations under 1 trait or 2 traits without pleiotropy, the univariate effect size of a new QTN mutation was drawn from a normal distribution with a mean of 0 and standard deviation sigma QTN. For QTN mutations under 2 traits with pleiotropy, the bivariate effect size was drawn from a multivariate normal distribution with a standard deviation of sigma QTN for both traits and no covariance, which gave flexibility for mutations to evolve with effects on one or both traits. Thus, the distribution of effect sizes and linkage relationships among QTNs was allowed to evolve.

Pleiotropy: Within each genic category were 5 levels of pleiotropy and selection: (i) 1 temperature trait (which adapted to the latitudinal cline), (ii) 2 traits without pleiotropy and equal strengths of selection on both traits, (iii) 2 traits without pleiotropy and with weaker selection on the temperature trait, (iv) 2 traits with pleiotropy (QTNs could evolve effects on one or both traits) and equal strengths of selection on both traits, and (v) 2 traits with pleiotropy and with weaker selection on the latitudinal temperature trait.

Selection: The trait was subject to spatially heterogeneous stabilizing selection with the optimum for each location in space given by the environment. For each individual in each generation, the fitness was determined by a Gaussian function given the difference between the individual's phenotype and the optimum at that location.

See trait equation details in the supplemental document:
equation_details_genetic_architecture_and_stabilizing_selection.pdf

For information on burn-in, adding neutral loci with tree sequencing, filtering, and sampling, see the Supplemental Methods.

Quantifying the degree of local adaptation, divergence, and structure

For each replicate, the degree of local adaptation was measured as (i) the difference between population fitness in sympatry and allopatry following (Blanquart et al. 2013) and (ii) the correlation between the phenotype and environmental cline for each trait. Overall divergence (genetic differentiation) was calculated as Weir and Cockerham's FST (Weir & Cockerham 1984) in OutFLANK (Whitlock & Lotterhos 2015). Population structure was estimated with a principal component analysis on the genotype matrix.

Quantifying trait and allelic clines

The degree of a trait cline was measured as Kendall's tau rank correlation coefficient between individual trait values and deme environment. The degree of an allele frequency cline was measured as Kendall's tau rank correlation coefficient (Kendall 1938, 1945) between deme allele frequency and deme environment, with significance being determined after Bonferroni correction based on the number of SNPs in the data. QTNs that were significant by this criteria were deemed “clinal QTNs.” The proportion of clinal QTNs excluded minor alleles with frequency < 0.01.

GEA performance

Latent-factor mixed models (LFMM) assess the linear relationship between genotype and environment while controlling for structure as latent factors. LFMM was implemented using the function lfmm2 in the R package LEA v.4.0.3 (Frichot et al. 2013; Frichot & François 2015; Caye et al. 2019). Redundancy analysis (RDA) and the partial RDA (pRDA) including a structure correction (conditional on the first 2 PC axes) were implemented using the ‘rda’ function in the R package vegan (Dixon 2003). See Supplemental Methods for details of implementation and correction for false discovery rate.

The performance of the association metrics were summarized as: (i) false discovery rate (FDR, proportion of outliers that are neutral, lower is better), (ii) true positive rate (TPR, proportion of QTNs that are significant outliers, higher is better), and (iii) the area under the precision-recall curve (AUC-PR, higher is better) (Lotterhos et al. 2022). In order to provide the most optimistic estimate of a method’s performance, the performance statistics were calculated by only including truly neutral loci unaffected by selection on linkage groups 11-20 and the QTNs.

Importance of clinal QTNs to local adaptation

First framework: A linear model was used to conduct the GWAS with individual trait value as the response variable and SNP genotype, PC1, and PC2 as explanatory variables. The proportion of GWAS hits that also showed clines with the environmental variable was compared to the known number of clinal QTNs. Second framework: The proportion of additive genetic variance (VA) for each QTN was calculated as the additive genetic variance for the focal QTN standardized by the total additive genetic variance following (Lotterhos 2019). The proportion of additive genetic variance (VA) explained by clinal QTNs was compared to a null expectation equal to the proportion of QTNs that were clinal. Third framework: I estimated the proportion of local explained by different subsets of QTNs: (i) QTNs with MAF > 0.01, (ii) clinal QTNs, and (iii) clinal QTNs inferred from latent factor mixed models that include a structure correction. For (ii) and (iii), a GEA model was performed for each environment, and then outlier QTNs were combined into a focal QTN set that was used for the local adaptation prediction. For each focal subset of QTNs, the counts of the derived allele were multiplied by the QTN effect size, summed to get a phenotype, and that phenotype was used in an in silico reciprocal transplant using the known phenotype-fitness function to estimate the degree of local adaptation. This estimate was then divided by the total degree of local adaptation (using all QTNs including those below the MAF threshold) to get an estimate of the proportion of local adaptation explained by that focal subset.

Data Processing Description

Processing notes from researcher:

- Conda environments used to create and process the simulations are included in the code repository

[table of contents | back to top]

Data Files

File
Summary data table for dataset 889769 filename: summary_20220428_20220726.csv (Comma Separated Values (.csv), 2.47 MB) MD5:8bf218995df195620cd3a1462e7bb324 File processed with laminar pipeline "889769_v1_paradox_of_adaptive_trait_clines" at path 889769/1/data/summary_20220428_20220726.csv

[table of contents | back to top]

Supplemental Files

File
Genotypes filename: genotypes.tar (Tape Archive (.tar), 8.87 GB) MD5:f0ea2afc1c65cb3b466dac668259c89a For each simulation seed, a genotype matrix with SNPs in rows and individuals in columns. 1000 individuals were sampled from the landscape (10/deme) and SNPs were filtered to MAF > 0.01. Each entry in the matrix is a 0, 1, or 2 corresponding to the counts of the derived allele.
Individual metadata filename: seed_Rout_ind_subset.txt_metadata.md (Plain Text, 3.38 KB) MD5:d90914a5c7eda47b3b54b57af8f955a2 Metadata describing the columns in the "Individuals" file
Individuals compressed file filename: individuals.tar (Tape Archive (.tar), 372.71 MB) MD5:c936fac46abfbcf224f3d31be75fb894 Data for each of the 1000 sampled individuals, for each of the 2250 simulation seeds
Mutations filename: mutations.tar (Tape Archive (.tar), 14.21 GB) MD5:260e2f85de28a462de748e81bf62366a Data for each SNP mutation in a simulation seed. Data correspond to rows in the Genotypes file.
Mutations metadata filename: seed_Rout_muts_full.txt_metadata.md (Plain Text, 7.76 KB) MD5:76b68fe8451b5895268332128428b3f6 Metadata for the mutation data describing each column in the data
Summary file filename: summary_20220428_20220726.txt (Plain Text, 3.63 MB) MD5:06ce383f26faeac78bbb7ab8e29489c3 Summary statistics for each simulation seed
Summary file metadata filename: summary_20220428_20220726.txt_metadata.md (Plain Text, 22.78 KB) MD5:470a7d6e497bc3e5046ef7409f88df26 Metadata for the summary file describing each column in the data

[table of contents | back to top]

Related Publications

Blanquart, F., Kaltz, O., Nuismer, S. L., & Gandon, S. (2013). A practical guide to measuring local adaptation. Ecology Letters, 16(9), 1195–1205. https://doi.org/10.1111/ele.12150
Methods

Coop, G., Witonsky, D., Di Rienzo, A., & Pritchard, J. K. (2010). Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Genetics, 185(4), 1411–1423. https://doi.org/10.1534/genetics.110.114819
Methods

Dixon, P. (2003). VEGAN, a package of R functions for community ecology. Journal of Vegetation Science, 14(6), 927–930. doi:10.1111/j.1654-1103.2003.tb02228.x
Methods

Forester, B. R., Jones, M. R., Joost, S., Landguth, E. L., & Lasky, J. R. (2015). Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. Molecular Ecology, 25(1), 104–120. Portico. https://doi.org/10.1111/mec.13476
Methods

Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype-environment associations. Molecular Ecology, 27(9), 2215–2233. Portico. https://doi.org/10.1111/mec.14584
Methods

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution, 6(8), 925–929. Portico. https://doi.org/10.1111/2041-210x.12382
https://doi.org/10.1111/2041-210X.12382
Methods

Frichot, E., Schoville, S. D., Bouchard, G., & François, O. (2013). Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. Molecular Biology and Evolution, 30(7), 1687–1699. https://doi.org/10.1093/molbev/mst063
Methods

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

Günther, T., & Coop, G. (2013). Robust Identification of Local Adaptation from Allele Frequencies. Genetics, 195(1), 205–220. https://doi.org/10.1534/genetics.113.152462
Methods

KENDALL, M. G. (1938). A NEW MEASURE OF RANK CORRELATION. Biometrika, 30(1–2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81
Methods

KENDALL, M. G. (1945). THE TREATMENT OF TIES IN RANKING PROBLEMS. Biometrika, 33(3), 239–251. https://doi.org/10.1093/biomet/33.3.239
Methods

Lotterhos, K. E. (2019). The Effect of Neutral Recombination Variation on Genome Scans for Selection. G3 Genes|Genomes|Genetics, 9(6), 1851–1867. https://doi.org/10.1534/g3.119.400088
Methods

Lotterhos, K. E., & Whitlock, M. C. (2015). The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Molecular Ecology, 24(5), 1031–1046. Portico. https://doi.org/10.1111/mec.13100
Methods

Lotterhos, K. E., Fitzpatrick, M. C., & Blackmon, H. (2022). Simulation Tests of Methods in Evolution, Ecology, and Systematics: Pitfalls, Progress, and Principles. Annual Review of Ecology, Evolution, and Systematics, 53(1), 113–136. https://doi.org/10.1146/annurev-ecolsys-102320-093722
Methods

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. Molecular Ecology, 24(17), 4348–4370. Portico. https://doi.org/10.1111/mec.13322
Methods

Weir, B. S., & Cockerham, C. C. (1984). ESTIMATING F -STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. Evolution, 38(6), 1358–1370. Portico. https://doi.org/10.1111/j.1558-5646.1984.tb05657.x
Methods

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution ofFST. The American Naturalist, 186(S1), S24–S36. https://doi.org/10.1086/682949
Methods

de Villemereuil, P., Frichot, É., Bazin, É., François, O., & Gaggiotti, O. E. (2014). Genome scan methods against more complex models: when and how much should we trust them? Molecular Ecology, 23(8), 2006–2019. https://doi.org/10.1111/mec.12705
Methods

[table of contents | back to top]

Parameters

Parameter	Description	Units
seed	Simulation seed.	unitless
n_samp_tot	Total number of individuals sampled.	unitless
n_samp_per_pop	Number of individuals sampled from each deme.	unitless
sd_fitness_among_inds	Variance in fitness among all sampled individuals in the simulation (sampling prob. is proportional to fitness to mimic viability selection).	unitless

sd_fitness_among_pops	Variance in fitness among all demes in the simulation after sampling (sampling prob. is proportional to fitness to mimic viability selection).	unitless
final_LA	Final amount of local adaptation in the simulation.	unitless
K	Number of populations used in analyses.	unitless
Bonf_alpha	The significance threshold for P-values applied to the correlation.	unitless
numCausalLowMAFsample	Number of causal loci that were not filtered out, but were below the MAF cutoff. These were included in the calculations.	unitless
all_corr_phen_temp	For all individuals, correlation between individual temp phenotype and environment temperature.	unitless
subsamp_corr_phen_temp	After sampling 10 individuals from each deme with a probability based on their fitness, correlation individual temp phenotype and environment temperature.	unitless
all_corr_phen_sal	For all individuals, correlation between individual sal phenotype and environment salinity.	unitless
subsamp_corr_phen_sal	After sampling 10 individuals from each deme with a probability based on their fitness, correlation between individual sal phenotype and environment salinity.	unitless
num_causal_prefilter	Number of causal loci in sim before filtering for MAF > 0.01.	unitless
num_causal_postfilter	Number of causal loci in sim before filtering for MAF > 0.01.	unitless
num_non_causal	Number of neutral loci in sim arising on the half of the genome where they could be linked to causal loci.	unitless
num_neut_prefilter	Total number of neutral loci on all LG before filtering MAF > 0.01. This is not really accurate, since many neutral loci were filtered out after output by pyslim - causal loci were not subject to filtering.	unitless
num_neut_postfilter	Total number of neutral loci on all LG after filtering MAF > 0.01.	unitless
num_neut_neutralgenome	Number of truly neutral loci in sim, unlinked to causal loci, on LG 11-20.	unitless
num_causal_temp	Number of loci with non-zero phenotypic effects on the temperature phenotype.	unitless
num_causal_sal	Number of loci with non-zero phenotypic effects on the salinity phenotype.	unitless
num_multiallelic	Rarely there is a back-mutation in SLiM, leading to a 0/0 0/1 1/1 1/2 0/2 2/2 genotypes. These were filtered for analysis.	unitless
meanFst	Overall FST (fixation index) calculated from mean(T1)/mean(T2) in outflank.	unitless
va_temp_total	Total additive genetic variance in the temperature trait, based on the entire 10,000 individual sample.	unitless
va_sal_total	Total additive genetic variance in the salinity trait, based on the entire 10,000 individual sample.	unitless
Va_temp_sample	Total additive genetic variance in the temperature trait, based on the 1,000 individual (10 ind/deme x 100 demes) sample.	unitless

Va_sal_sample	Total additive genetic variance in the salinity trait, based on the entire 1,000 individual (10 ind/deme x 100 demes) sample.	unitless
nSNPs	Total number of SNPs in analysis.	unitless
median_causal_temp_cor	Median abs(Spearman's correlation) between allele frequency and temperature for causal loci.	unitless
median_causal_sal_cor	Median abs(Spearman's correlation) between allele frequency and salinity for causal loci.	unitless
median_neut_temp_cor	Median abs(Spearman's correlation) between allele frequency and temperature for neutral loci.	unitless
median_neut_sal_cor	Median abs(Spearman's correlation) between allele frequency and salinity for neutral loci.	unitless
cor_VA_temp_prop	Proportion of VA in temperature phenotype explained by clinal outliers for temperature, based on kendall's correlation between deme allele frequency and deme temperature.	unitless
cor_VA_sal_prop	Proportion of VA in salinity phenotype explained by clinal outliers for salinity, based on kendall's correlation between deme allele frequency and deme salinity.	unitless
cor_TPR_temp	True positive rate for loci with non-zero effects on temperature, based on kendall's correlation between deme allele frequency and deme temperature.	unitless
cor_TPR_sal	True positive rate for loci with non-zero effects on salinity, based on kendall's correlation between deme allele frequency and deme salinity.	unitless
cor_FDR_allSNPs_temp	False discovery rate of (kendall's correlation between deme allele frequency and deme temperature) for loci with non-zero effects on temperature.	unitless
cor_FDR_neutSNPs_temp	An optimistic calculation for false discovery rate of (kendall's correlation between deme allele frequency and deme temperature) for loci with non-zero effects on temperature, excluding non-causal loci in half of genome affected by selection.	unitless
cor_FDR_allSNPs_sal	False discovery rate of (kendall's correlation between deme allele frequency and deme salinity) for loci with non-zero effects on salinity.	unitless
cor_FDR_neutSNPs_sal	An optimistic calculation for false discovery rate of (kendall's correlation between deme allele frequency and deme temperature) for loci with non-zero effects on temperature, excluding non-causal loci in half of genome affected by selection.	unitless
num_causal_sig_temp_corr	Number of causal loci on temperature trait that are significant cor(af,temp) after Bonferroni correction.	unitless
num_causal_sig_sal_corr	Number of causal loci on salinity trait that are significant cor(af,salinity) after Bonferroni correction.	unitless
num_notCausal_sig_temp_corr	Number of non-causal (neutral and neutral-linked) loci that are significant cor(af,temp) after Bonferroni correction.	unitless
num_notCausal_sig_sal_corr	Number of non-causal (neutral and neutral-linked) loci that are significant cor(af,salinity) after Bonferroni correction.	unitless
num_neut_sig_temp_corr	Number of truly neutral loci (LG 11-20) that are significant cor(af,temp) after Bonferroni correction.	unitless

num_neut_sig_sal_corr	Number of truly neutral loci (LG 11-20) that are significant cor(af,salinity) after Bonferroni correction.	unitless
cor_AUCPR_temp_allSNPs	AUC-PR of kendall's correlation between deme allele frequency and deme temperature, based on the whole genome and causal loci for temperature.	unitless
cor_AUCPR_temp_neutSNPs	An optimistic estimate of AUC-PR of kendall's correlation between deme allele frequency and deme temperature, based on causal loci for temperature and neutral loci not affected by selection (excluding non-causal loci in half of genome affected by selection).	unitless
cor_AUCPR_sal_allSNPs	AUC-PR of kendall's correlation between deme allele frequency and deme salinity, based on the whole genome and causal loci for salinity.	unitless
cor_AUCPR_sal_neutSNPs	An optimistic estimate of AUC-PR of kendall's correlation between deme allele frequency and deme salinity, based on causal loci for salinity and neutral loci not affected by selection (excluding non-causal loci in half of genome affected by selection).	unitless
cor_af_temp_noutliers	Number of outliers for cor(af,temp) after Bonferroni correction.	unitless
cor_af_sal_noutliers	Number of outliers for cor(af,salinity) after Bonferroni correction.	unitless
cor_FPR_temp_neutSNPs	False positive rate in cor(af,temp) after Bonferroni correction, based on neutral loci unaffected by selection (LG 11-20).	unitless
cor_FPR_sal_neutSNPs	False positive rate in cor(af,sal) after Bonferroni correction, based on neutral loci unaffected by selection (LG 11-20).	unitless
LEA3_2_lfmm2_Va_temp_prop	Proportion of additive genetic variance (Va) in the temperature trait explained by outliers in the LFMM temp model.	unitless
LEA3_2_lfmm2_Va_sal_prop	Proportion of additive genetic variance (Va) in the saliniity trait explained by outliers in the LFMM salinity model.	unitless
LEA3_2_lfmm2_TPR_temp	True positive rate of the LFMM temp model for loci with alleles that have non-zero effects on the temperature phenotype.	unitless
LEA3_2_lfmm2_TPR_sal	True positive rate of the LFMM salinity model for loci with alleles that have non-zero effects on the salnity phenotype.	unitless
LEA3_2_lfmm2_FDR_allSNPs_temp	False discovery rate of the LFMM temp model for the entire genome.	unitless
LEA3_2_lfmm2_FDR_allSNPs_sal	False discovery rate of the LFMM temp model for the entire genome.	unitless
LEA3_2_lfmm2_FDR_neutSNPs_temp	An optimistic calculation of the false discovery rate of the LFMM temp model, including only causal loci and neutral loci not affected by selection (any non-causal loci that arises on the half of the genome affected by selection was excluded).	unitless
LEA3_2_lfmm2_FDR_neutSNPs_sal	An optimistic calculation of the false discovery rate of the LFMM salinity model, including only causal loci and neutral loci not affected by selection (any non-causal loci that arises on the half of the genome affected by selection was excluded).	unitless

LEA3_2_lfmm2_AUCPR_temp_allSNPs	The AUC-PR of the lfmm temp model based on the entire genome.	unitless
LEA3_2_lfmm2_AUCPR_temp_neutSNPs	An optimistic calculation of the AUC-PR of the LFMM temp model, including only causal loci and neutral loci not affected by selection (any non-causal loci that arises on the half of the genome affected by selection was excluded).	unitless
LEA3_2_lfmm2_AUCPR_sal_allSNPs	The AUC-PR of the lfmm salinity model based on the entire genome.	unitless
LEA3_2_lfmm2_AUCPR_sal_neutSNPs	An optimistic calculation of the AUC-PR of the LFMM salinity model, including only causal loci and neutral loci not affected by selection (any non-causal loci that arises on the half of the genome affected by selection was excluded).	unitless
LEA3_2_lfmm2_mlog10P_tempenv_noutliers	Number of outliers for the lfmm temp model (qvalue	unitless
LEA3_2_lfmm2_mlog10P_salenv_noutliers	Number of outliers for the lfmm salinity model (qvalue	unitless
LEA3_2_lfmm2_num_causal_sig_temp	Number of causal loci on the temp trait, significant in the lfmm temp model (qvalue	unitless
LEA3_2_lfmm2_num_neut_sig_temp	Number of neutral loci false positives (only neutral loci not affected by selection), significant in the lfmm temp model (qvalue	unitless
LEA3_2_lfmm2_num_causal_sig_sal	Number of causal loci on the salinity trait, significant in the lfmm salinity model (qvalue	unitless
LEA3_2_lfmm2_num_neut_sig_sal	Number of neutral loci false positives (only neutral loci not affected by selection), significant in the lfmm salinity model (qvalue	unitless
LEA3_2_lfmm2_FPR_neutSNPs_temp	False positive rate of lfmm temperature model.	unitless
LEA3_2_lfmm2_FPR_neutSNPs_sal	False positive rate of lfmm salinity model.	unitless
RDA1_propvar	Proportion of variance explained by first RDA axis. RDA model: genotype ~ environment.	unitless
RDA2_propvar	Proportion of variance explained by second RDA axis. RDA model: genotype ~ environment.	unitless
RDA1_propvar_corr	Proportion of variance explained by first RDA axis. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA2_propvar_corr	Proportion of variance explained by second RDA axis. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA1_temp_cor	Output of `summary(rdaout)$biplot[2,1]`, which is the correlation between RDA1 and the temperature environmental variable. RDA model: genotype ~ environment.	unitless
RDA1_sal_cor	Output of `summary(rdaout)$biplot[1,1]`, which is the correlation between RDA1 and the salinity environmental variable. RDA model: genotype ~ environment.	unitless
RDA2_temp_cor	Output of `summary(rdaout)$biplot[2,2]`, which is the correlation between RDA2 and the temperature environmental variable. RDA model: genotype ~ environment.	unitless
RDA2_sal_cor	Output of `summary(rdaout)$biplot[1,2]`, which is the correlation between RDA2 and the salinity environmental variable. RDA model: genotype ~ environment.	unitless

RDA_Va_temp_prop	Proportion of additive genetic variance (Va) in the temperature trait explained by outliers in the RDA outlier analysis, following (Capblanq 2018, https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12906). RDA model: genotype ~ environment.	unitless
RDA_Va_temp_prop_corr	Proportion of additive genetic variance (Va) in the temperature trait explained by outliers in the RDA outlier analysis, following (Capblanq 2018, https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12906). RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_Va_sal_prop	Proportion of additive genetic variance (Va) in the salinity trait explained by outliers in the RDA outlier analysis, following (Capblanq 2018, https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12906). RDA model: genotype ~ environment.	unitless
RDA_Va_sal_prop_corr	Proportion of additive genetic variance (Va) in the salinity trait explained by outliers in the RDA outlier analysis, following (Capblanq 2018, https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.12906). RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_TPR	True positive rate of the RDA for all causal loci. Since RDA is a multidimensional analysis, I did not differentiate between loci that had causal effects on temperature or salinity. RDA model: genotype ~ environment.	unitless
RDA_TPR_corr	True positive rate of the RDA for all causal loci. Since RDA is a multidimensional analysis, I did not differentiate between loci that had causal effects on temperature or salinity. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_FDR_allSNPs	False discovery rate of the RDA outlier analysis based on the entire genome. RDA model: genotype ~ environment.	unitless
RDA_FDR_allSNPs_corr	False discovery rate of the RDA outlier analysis based on the entire genome. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
num_RDA_sig_causal	Number of causal loci that are significant in the RDA analysis at q > 0.05. RDA model: genotype ~ environment.	unitless
num_RDA_sig_neutral	Number of neutral loci (LG 11-20) that are significant in the RDA analysis at q > 0.05. RDA model: genotype ~ environment.	unitless
num_RDA_sig_causal_corr	Number of causal loci that are significant in the RDA analysis at q > 0.05. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
num_RDA_sig_neutral_corr	Number of neutral loci (LG 11-20) that are significant in the RDA analysis at q > 0.05. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless

RDA_FDR_neutSNPs	An optimistic calculation of the false discovery rate of the RDA, including only causal loci and neutral loci not affected by selection (any non-causal loci that arises on the half of the genome affected by selection was excluded). RDA model: genotype ~ environment.	unitless
RDA_FDR_neutSNPs_corr	An optimistic calculation of the false discovery rate of the RDA, including only causal loci and neutral loci not affected by selection (any non-causal loci that arises on the half of the genome affected by selection was excluded). RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_AUCPR_allSNPs	AUC-PR of the RDA outlier analysis based on the entire genome. RDA model: genotype ~ environment.	unitless
RDA_AUCPR_neutSNPs	An optimistic calculation of the AUC-PR of the RDA, including only causal loci and neutral loci not affected by selection (any non-causal loci that arises on the half of the genome affected by selection was excluded). RDA model: genotype ~ environment.	unitless
RDA_AUCPR_neutSNPs_corr	An optimistic calculation of the AUC-PR of the RDA, including only causal loci and neutral loci not affected by selection (any non-causal loci that arises on the half of the genome affected by selection was excluded). RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_FPR_neutSNPs	False positive rate of the RDA analysis, based only on neutral SNPs. RDA model: genotype ~ environment.	unitless
RDA_FPR_neutSNPs_corr	False positive rate of the RDA analysis, based only on neutral SNPs. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_RDAmutpred_cor_tempEffect	# pearson's correlation between the predicted temperature effect from RDA and the true mutation effect on temperature. RDA model: genotype ~ environment.	unitless
RDA_RDAmutpred_cor_salEffect	# pearson's correlation between the predicted salinity effect from RDA and the true mutation effect on salinity. RDA model: genotype ~ environment.	unitless
RDA_absRDAmutpred_cor_tempVa	# pearson's correlation between the abs(predicted temperature effect from RDA) and the true mutation Va on temperature. RDA model: genotype ~ environment.	unitless
RDA_absRDAmutpred_cor_salVa	# pearson's correlation between the abs(predicted salinity effect from RDA) and the true mutation Va on salinity. RDA model: genotype ~ environment.	unitless
RDA_RDAmutpred_cor_tempEffect_structcorr	# pearson's correlation between the predicted temperature effect from RDA and the true mutation effect on temperature. RDA model: genotype ~ environment. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_RDAmutpred_cor_salEffect_structcorr	# pearson's correlation between the predicted salinity effect from RDA and the true mutation effect on salinity. RDA model: genotype ~ environment. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless

RDA_absRDAmutpred_cor_tempVa_structcorr	# pearson's correlation between the abs(predicted temperature effect from RDA) and the true mutation Va on temperature. RDA model: genotype ~ environment. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_absRDAmutpred_cor_salVa_structcorr	# pearson's correlation between the abs(predicted salinity effect from RDA) and the true mutation Va on salinity. RDA model: genotype ~ environment. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_cor_RDA20000temppredict_tempPhen	Correlation between the true temperature phenotype and that predicted from an RDA based on 20K SNPs. see `seed_Rout_RDA_predictions` for correlations with less loci used to make the prediction. RDA model: genotype ~ environment.	unitless
RDA_cor_RDA20000salpredict_salPhen	Correlation between the true salinity phenotype and that predicted from an RDA based on 20K SNPs. see `seed_Rout_RDA_predictions` for correlations with less loci used to make the prediction. RDA model: genotype ~ environment.	unitless
RDA_cor_RDA20000temppredict_tempPhen_structcorr	Correlation between the true temperature phenotype and that predicted from an RDA based on 20K SNPs. see `seed_Rout_RDA_predictions` for correlations with less loci used to make the prediction. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
RDA_cor_RDA20000salpredict_salPhen_structcorr	Correlation between the true salinity phenotype and that predicted from an RDA based on 20K SNPs. see `seed_Rout_RDA_predictions` for correlations with less loci used to make the prediction. RDA model with structure correction: genotype ~ environment + Condition(PC1 + PC2).	unitless
cor_PC1_temp	Correlation between individual loading on PC1 from the principle components based on the Genotype-matrix (individual genotypes labeled as 0,1,2) and temperature of the deme where it was sampled.	unitless
cor_PC1_sal	Correlation between individual loading on PC1 from the principle components based on the Genotype-matrix and salnity of the deme where it was sampled.	unitless
cor_PC2_temp	Correlation between individual loading on PC2 from the principle components based on the Genotype-matrix and temperature of the deme where it was sampled.	unitless
cor_PC2_sal	Correlation between individual loading on PC2 from the principle components based on the Genotype-matrix and salnity of the deme where it was sampled.	unitless
cor_LFMMU1_temp	Correlation between the individual loading on the latent factor 1 from the lfmm model based on temperature.	unitless
cor_LFMMU1_sal	Correlation between the individual loading on the latent factor 1 from the lfmm model based on salnity.	unitless
cor_LFMMU2_temp	Correlation between the individual loading on the latent factor 2 from the lfmm model based on temperature.	unitless

cor_LFMMU2_sal	Correlation between the individual loading on the latent factor 2 from the lfmm model based on salnity.	unitless
cor_PC1_LFMMU1_temp	Correlation between (individual loading on PC1 from the principle components based on the Genotype-matrix) and (individual loading on the latent factor 1 from the lfmm model based on temperature).	unitless
cor_PC1_LFMMU1_sal	Correlation between (individual loading on PC1 from the principle components based on the Genotype-matrix) and (individual loading on the latent factor 1 from the lfmm model based on salinity).	unitless
cor_PC2_LFMMU1_temp	Correlation between (individual loading on PC2 from the principle components based on the Genotype-matrix) and (individual loading on the latent factor 1 from the lfmm model based on temperature).	unitless
cor_PC2_LFMMU1_sal	Correlation between (individual loading on PC2 from the principle components based on the Genotype-matrix) and (individual loading on the latent factor 1 from the lfmm model based on salinity).	unitless
gwas_TPR_sal	True positive rate of the GWAS model for the salinity trait.	unitless
gwas_TPR_temp	True positive rate of the GWAS model for the temperature trait.	unitless
gwas_FDR_sal_neutbase	False discovery rate of the GWAS model for the salinity trait, only including QTNs and purely neutral loci unaffected by selection.	unitless
gwas_FDR_temp_neutbase	False discovery rate of the GWAS model for the temperature trait, only including QTNs and purely neutral loci unaffected by selection.	unitless
clinalparadigm_sal_proptop5GWASclines	Proportion of the top 5% of GWAS loci with the smallest P-values for the salinity trait (true and false positives) that show clines.	unitless
clinalparadigm_temp_proptop5GWASclines	Proportion of the top 5% of GWAS loci with the smallest P-values for the temperature trait (true and false positives) that show clines.	unitless
clinalparadigm_sal_propsigGWASclines	Proportion of GWAS hits for the salinity trait (true and false positives) that show clines.	unitless
clinalparadigm_temp_propsigGWASclines	Proportion of GWAS hits for the temperature trait (true and false positives) that show clines.	unitless

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Northeastern's High Performance Computing Cluster
Generic Instrument Name	High-Performance Computing Cluster
Generic Instrument Description	"High-Performance Computing" (HPC) refers to a class of evolving technologies that provide leading-edge computational capabilities, including scalable high-performance computers, high-end graphic systems, and high-speed networks. HPC may be used for molecular modeling, genome analysis, and image processing, among others.

[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]