Population genomics (RADseq-derived SNP genotypes) of the amphipod Ampithoe valida from collections in Japan, North America, and South America in 2015

Website: https://www.bco-dmo.org/dataset/998972
Data Type: Other Field Results
Version: 1
Version Date: 2026-05-19

Project
» Detecting genetic adaptation during marine invasions (Genetic Adaptation Marine Inv)
» The genetic legacy of an Asian oyster introduction and its disease-causing parasite (Oyster historical genetics)

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

Abstract
We clarified the cryptic diversity and introduction history of the marine amphipod Ampithoe valida by genotyping 10,295 single-nucleotide polymorphisms (SNPs) for 349 individuals from Japan, North America and Argentina. This dataset includes metadata, methods, genetic accession identifiers, and links to published code and processed data used to generate figures for the results publication Harper et al. (2022, doi:10.1007/s10592-022-01452-8) titled "Global distribution of cryptic native, introduced and hybrid lineages in the widespread estuarine amphipod Ampithoe valida." Raw FASTQ files were contributed to the National Center for Biotechnology Information (NCBI)'s Sequence Read Archive (SRA) and are available under BioProject PRJNA825556.

Table of Contents

Coverage

Location: intertidal and shallow subtidal (<1m MLLW). Worldwide estuaries.
Spatial Extent: **N**:46.66667 **E**:142.5 **S**:-43.37 **W**:-124.216667
Temporal Extent: 2015

Methods & Sampling

Individuals were collected by hand from eelgrass beds and from algae (primarily *Ulva lactuca*, *Gracilaria vermiculophylla* and *Fucus vesiculosus)* on docks, mudflats, and sandy/rocky shores. They were immediately euthanized and stored in 95% ethanol until use. Genomic DNA was extracted from amphipod tissue using either a Qiagen DNeasy spin-column kit following the manufacturer’s protocol for animal tissues or a Macherey-Nagel Nucleospin Tissue kit. For individuals that were less than approximately one centimeter in

length or severely degraded, the entire individual was used; otherwise only half of the individual was used in the extraction. We genotyped single nucleotide polymorphisms (SNPs) using a restriction-site associated genotype-by-sequencing on DNA extractions from 350 individuals. Eleven individuals sampled for mtDNA, but not included in the RADseq library did not have a sufficient amount of DNA extraction product to be included. Genomic DNA was digested with two restriction enzymes, EcoRI and MseI. Pooled fragments for each individual were ligated with customized adaptor sequences containing the Illumina adaptors and primer sequences and unique 8–10 bp barcodes to allow for the in silico identification of individuals. DNA fragments were amplified by PCR twice and size selected (300–450 bp) using a Blue Pippin. The final library was sequenced using a single-end 100 bp protocol within a lane of an Illumina HiSeq 2500 sequencer at the University of Texas at Austin Genome Sequencing and Analysis Facility (GSAF).

Organism name, Life Science Identifier (LSID):
Ampithoe valida, urn:lsid:marinespecies.org:taxname:102005

Data Processing Description

PhiX sequences present in the raw reads were identified and removed (13% was phiX). Cut sites and adaptors were removed from the remaining reads (256,667,215) and then parsed using barcodes into a single FASTQ file per individual using custom *Perl* scripts. A subset of 15 million raw reads was used to create a de novo assembly. Contigs were assembled with a minimum match percent set to 92% and pruned with a minimum sequence length of 50 bp. To remove potential paralogs, consensus sequences obtained for each contig were assembled de novo to each other with a minimum match percent set to 83%. Parsed reads were then assembled, with a maximum edit distance set to 6 bp, to the pruned contigs coming out of the de novo assembly. These alignments were then used to call variant SNP sites. We analyzed a single SNP per contig to reduce the effects of physical linkage on population genetic parameters and filtered out SNPs that were recorded in fewer than 50% of individuals, had greater than two alleles per individual and possessed a minor allele frequency < 5%. We generated a set of genotype likelihoods that combine the uncertainty generated by sequence coverage, sequencing error, and alignment error. We converted the phred-scale genotype likelihoods (from samtools/bcftools) per SNP-sample combination into probabilities that summed to 1, and then converted these to a single value that ranges from 0 to 2, where 0, 1 and 2 represent the highest probability of a homozygote, heterozygote, and alternative homozygote, respectively.

BCO-DMO Processing Description

- Loaded "attributes_final.csv" as table "998972_v1_a-valida-pop-genomics" (CSV format, header row 1), treating empty strings and "nd" as missing values
- Applied metadata (descriptions, standard name IDs, units) to columns: Accession, BioSample, collection_date, geo_loc_name, isolate, sample_name, strain
- Split column "lat_lon" (format: decimal degrees + direction, e.g. "12.345 N 67.890 W") into lat,lon column
- Set lat and lon set to number type
- Deleted original lat_lon column after verifying coordinate parsing was correct
- Rounded latitude to 5 decimal places. The coordinates were provided with varying levels of precision from degree to degree with several decimal places. Integers were not padded with decimal 0000s.
- Added constant column "BioProject" with value "PRJNA825556"
- Reordered columns to: BioProject, BioSample, sample_name, strain, isolate, collection_date, geo_loc_name, Accession, latitude, longitude
- Updated metadata (descriptions, standard name IDs, units) for all columns, including latitude and longitude marked as primary parameters with units of decimal degrees
- Output saved as "998972_v1_a-valida-pop-genomics.csv"

Problem Description

NA

[table of contents | back to top]

Related Publications

Harper, K. E., Scheinberg, L. A., Boyer, K. E., & Sotka, E. E. (2022). Global distribution of cryptic native, introduced and hybrid lineages in the widespread estuarine amphipod Ampithoe valida. Conservation Genetics, 23(4), 791–806. https://doi.org/10.1007/s10592-022-01452-8
Results

[table of contents | back to top]

Related Datasets

IsRelatedTo

College of Charleston (2022). Global distribution of cryptic native, introduced and hybrid lineages in the widespread estuarine amphipod Ampithoe valida. 2022/04. NCBI:BioProject: PRJNA825556. In: BioProject [Internet]. Bethesda, MD: National Library of Medicine (US), National Center for Biotechnology Information. Available from: http://www.ncbi.nlm.nih.gov/bioproject/PRJNA825556

Erik Sotka. (2026). *esotka/AmpithoeValida: AmpithoeValidaPopGen* (Version v1.0) [Computer software]. Zenodo. https://doi.org/10.5281/ZENODO.18672806

[table of contents | back to top]

Parameters

Parameter	Description	Units
BioProject	National Center for Biotechnology Information (NCBI) BioProject identifier	unitless
BioSample	BioSample ID	unitless
sample_name	individual ID	unitless
strain	individual ID	unitless
isolate	Population ID	unitless
collection_date	Year of collection (yyyy)	unitless
geo_loc_name	Country	unitless
Accession	Sequence Read Archive (at NCBI) Run ID	unitless
latitude	latitude	decimal degrees
longitude	longitude	decimal degrees

[table of contents | back to top]

Instruments

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

Dataset-specific Instrument Name	96-well thermocycler (Eppendorf)
Generic Instrument Name	Thermal Cycler
Generic Instrument Description	A thermal cycler or "thermocycler" is a general term for a type of laboratory apparatus, commonly used for performing polymerase chain reaction (PCR), that is capable of repeatedly altering and maintaining specific temperatures for defined periods of time. The device has a thermal block with holes where tubes with the PCR reaction mixtures can be inserted. The cycler then raises and lowers the temperature of the block in discrete, pre-programmed steps. They can also be used to facilitate other temperature-sensitive reactions, including restriction enzyme digestion or rapid diagnostics. (adapted from http://serc.carleton.edu/microbelife/research_methods/genomics/pcr.html)

[table of contents | back to top]

Project Information

Detecting genetic adaptation during marine invasions (Genetic Adaptation Marine Inv)

Coverage: Estuaries of NW and NE Pacific; estuaries of NW and NE Atlantic

Description from NSF award abstract:
Biological introductions, defined as the establishment of species in geographic regions outside the reach of their natural dispersal mechanisms, have dramatically increased in frequency during the 20th century and are now altering community structure and ecosystem function of virtually all marine habitats. To date, studies on marine invasions focus principally on demographic and ecological processes, and the importance of evolutionary processes has been rarely tested. This knowledge gap has implications for management policies, which attempt to prevent biological introductions and mitigate their impacts. The Asian seaweed *Gracilaria vermiculophylla* has been introduced to every continental margin in the Northern Hemisphere, and preliminary data indicate that non-native populations are both more resistant to heat stress and resistant to snail herbivory. The project will integrate population genetics, field survey and common-garden laboratory experiments to comprehensively address the role of rapid evolutionary adaptation in the invasion success of this seaweed. Specifically, the PIs will answer the following. What is the consequence of introductions on seaweed demography and mating systems? How many successful introductions have occurred in North America and Europe? Where did introduced propagules originate? Do native, native-source and non-native locations differ in environmental conditions? Do native, native-source and non-native populations differ in phenotype?

The intellectual merit of this project is based on three gaps in the literature. First, while biological invasions are

widely recognized as a major component of global change, there are surprisingly few studies that compare native and non-native populations in their biology or ecology. Native and non-native populations will be surveyed in a similar manner, allowing assessment of differences in population dynamics, mating system, epifaunal and epiphytic communities, and the surrounding abiotic and biotic environment. Second, *G. vermiculophylla* exhibits a life cycle typical of other invasive species (including some benthic invertebrates), yet we still lack data on the effects of decoupling the haploid and diploid stages on genetic structure, and in turn, on the evolvability of their populations. Finally, this project will provide unequivocal evidence of an adaptive shift in a marine invasive. To our knowledge, such evolutionary change has been described previously for only a complex of marine copepod species. *G. vermiculophylla* will serve as a model for understanding evolution in other nuisance invasions, and perhaps lead to novel methods to counter future invasions or their spread.

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

Coverage: Global

NSF abstract:

During the 20th century, the Pacific oyster *Crassostrea gigas* was deliberately introduced from its native range of coastal Asia to the estuaries of six continents. While the introduced Pacific oysters are widely aquacultured and thus can generate local economic wealth, they sometimes outcompete native oysters, and can carry microbial, animal and plant hitchhikers that negatively impact local economies and the ecological functioning of local estuaries. This study comprehensively assesses the pathways and sources of Pacific oyster introductions using a worldwide, population genetic survey. Simultaneously, the study also assesses the pathways and source of one hitchhiking protist (*Haplosporidium nelsoni*) that causes the disease MSX (multinucleated sphere X) in the Virginia oyster (*Crassostrea virginica*) along the eastern seaboard of the United States. One goal of this research is to generate management strategies that combat the negative impacts of the Pacific oyster and its associated invaders, and minimize future invasions. A second goal is to minimize some uncertainty about the population biology of the devastating Haplosporidium parasite, and thus, increase confidence of policy makers who are managing shellfish health, restoration and commerce. By quantifying the pathways and sources of *C. gigas*, this project may inform strategies to combat negative impacts of *C. gigas* and its associated invaders, as well as minimize future invasions. Moreover, quantifying dispersal within and among populations of *H. nelsoni* along the US East Coast will provide perspective on the effectiveness of regional biosecurity measures in preventing the ongoing dispersal of this destructive pathogen via aquaculture. In addition, the project lends itself well to programs that foster critical thinking and research experience among both undergraduate and K-12 students. The project provides opportunities for 6-9 undergraduates to perform research, includes a 2-day workshop on bioinformatics for the wider undergraduate community, and facilitates ongoing opportunities for K-12 students to participate in citizen-science research.

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

This award reflects NSF's statutory mission and has been deemed worthy of support through evaluation using the Foundation's intellectual merit and broader impacts review criteria.

[table of contents | back to top]

Funding

Funding Source	Award
NSF Division of Ocean Sciences (NSF OCE)	OCE-1357386
NSF Division of Ocean Sciences (NSF OCE)	OCE-1924599

[table of contents | back to top]