Methodology information and links to data access for allele frequencies and FST estimates for 1,904,119 SNPs analyzed in five population samples of Atlantic silverside (Menidia menidia) collected along the east coast of North America between 2005 to 2007

Website: https://www.bco-dmo.org/dataset/854895
Data Type: Other Field Results
Version: 1
Version Date: 2021-06-29

Project
» Collaborative research: The genomic underpinnings of local adaptation despite gene flow along a coastal environmental cline (GenomAdapt)

Contributors	Affiliation	Role
Therkildsen, Nina Overgaard	Cornell University (Cornell)	Principal Investigator
Baumann, Hannes	University of Connecticut (UConn)	Co-Principal Investigator
York, Amber D.	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
Methodology information and links to data access for allele frequencies and FST estimates for 1,904,119 SNPs analyzed in five population samples of Atlantic silverside (Menidia menidia) across its distribution range. These data were published in Wilder et al. (2020). The allele frequencies and FST estimates and metadata are stored at the Dryad repository (https://doi.org/10.5061/dryad.jm63xsj7w).

Table of Contents

Coverage

Spatial Extent: **N**:47.4 **E**:-61.85 **S**:31.02 **W**:-81.43
Temporal Extent: 2005 - 2007

Methods & Sampling

Methodology:

The following is an excerpt of the methods used to generate these data. Full details can be found in the associated publication and supplemental material of Wilder et al. (2020) and is also available at the Dryad repository where the allele frequencies and FST estimates are stored (https://doi.org/10.5061/dryad.jm63xsj7w).

Data generation read mapping and SNP calling:

Atlantic silverside muscle samples collected from Georgia (GA), New York (NY), Gulf of Maine (GoM) and Gulf of St. Lawrence (GSL) were sequenced by Therkildsen & Palumbi (2017) to a final, average depth of 1.5x across the reference transcriptome. See Therkildsen & Palumbi (2017) and Therkildsen et al. (2019) for further detail. We added 47 additional samples from Oregon Inlet, North Carolina (NC) near Cape Hatteras. Individually barcoded sample libraries were prepared at Cornells Biotechnology Resource Center using similar methods to Therkildsen & Palumbi (2017), with a few minor modifications. Reagents from the Illumina Nextera kit (96 sample Nextera DNA Library Prep Kit) were used at 1/3 the recommended concentration in 1/10 the recommended volume (5ul instead of 50ul), with 2ng of input DNA. Individual libraries were pooled, and size-selected using a Pippin Prep to remove fragments <286 bp (150 bp insert plus 136 bp of Illumina adapters). Filtered mapped reads for the NC samples had a final average depth of 1.5x.

We generated paired-end 75 bp sequences on the NextSeq500. Filtered mapped reads for the NC samples had a final average depth of 1.5x.

Read adapters were trimmed from the paired-end reads using Trimmomatic v.0.3.6 (Bolger et al. 2014), and mapped to the reference transcriptome with Bowtie2 v2.2.3 (Langmead and Salzberg 2012), using the very-sensitive-local mode and retaining unpaired, orphaned and concordantly paired reads with mapping quality >20. Duplicate reads were removed with MarkDuplicates (PICARD Tools v1.139; http://broadinstitute.github.io/picard/). To avoid spurious SNP calls stemming from differences in read length in the NC sample (Leigh et al. 2018), we called SNPs across the original four population sample set (GA, NY, GoM and GSL). Downstream analyses of all samples (including NC) were performed on this set of SNPs. Because NC is located in the center of the distribution of the southern phylogeographic group (Mach et al. 2005; Lou et al. 2018), we do not expect to have missed variants private to that population by excluding it from the SNP calling. Biallelic SNPs were called in the program ANGSD v. 0.912 (Korneliussen et al. 2014). ANGSD estimates genotype likelihoods at each site based on the aligned reads and their mapping and sequencing quality scores. Using genotype likelihoods for low-coverage sequencing data allows for the incorporation of uncertainty in genotyping, providing more accurate population genetic inference (Li 2011; Nielsen et al. 2011). We called polymorphic sites across the four original populations (n = 189), at sites with a probability <1e-6 of being monomorphic, using the SAMtools model of genotype likelihoods in ANGSD. Bases with quality score <20 were excluded. We excluded sites with minor allele frequency (MAF) < 0.01, total read depth <75 and >759 (mean depth +1 standard deviation), or data from <75 individuals. ANGSD called 1,942,329 SNPs that passed the above filters. From these, we filtered out 38,210 SNPs within repetitive elements identified by RepeatMasker (Smit et al. 2017). The final dataset comprised 1,904,119 bialleleic SNPs across the ~52 MB transcriptome.

Estimating allele frequencies, identifying FST outliers and functional annotation of SNPs

At all globally variant sites, major and minor alleles were inferred based on global allele frequencies and minor allele frequencies (MAF) were estimated within populations from genotype likelihoods at sites with data for at least 10 individuals in the program ANGSD. We estimated pairwise FST at each SNP between population pairs using the 2-dimensional site frequency spectrum (2D SFS) as prior.

For each SNP, we estimated the global FST across all five populations as in the OutFLANK script FST functions.R, except that we made a slight modification to the script so that we could directly input allele frequencies and sample sizes estimated in ANGSD (which takes genotype likelihoods into account), rather than estimating allele frequencies from allele counts in OutFLANK. FST was then estimated by the procedure implemented in OutFLANK. We fit the X2 distribution to the pruned SNP set by trimming 5% from each side of the FST distribution and using q<0.05 to estimate the degrees of freedom and FSTbar of the X2 distribution in OutFLANK (Whitlock and Lotterhos 2015). We then applied the X2 fit to all SNPs across the transcriptome to identify FST outliers using q<0.05.

To understand the potential function of SNPs within the transcriptome, we used two programs, Transdecoder and GeneMarkS-T, to predict coding sequences (CDS) and untranslated regions (UTR). From these annotations, we used snpEff to predict the function (e.g., missense variant, synonymous variant, 3 UTR variant) of each SNP.

Identifying LD blocks and ordering Atlantic silverside genes along medaka chromosomes

We estimated linkage disequilibrium between all pairs of SNPs identified as FST outliers. Because the majority of these SNPs were fixed for opposite alleles at the extremes of the geographic distribution, precluding LD inference, we limited our LD analysis to the three sampling locations in the center of the range (NC, NY and GoM). Within each of these populations, we estimated LD across pairs of SNPs with MAF>0.1 in the program ngsLD. We summarized patterns by calculating an average LD between pairs of contigs that had at least 2

outlier SNPs with MAF>0.1. For each pair of these contigs, we calculated the mean D across all pairs of outlier SNPs between contigs. We used a hierarchical clustering approach to determine how contigs clustered into blocks within each population by generating dendrograms from the pairwise LD matrix using the Ward.D method in the hclust function in R (Murtagh and Legendre 2014).

We downloaded all medaka peptide sequences and used blastx to compare silversides contigs to medaka peptides. We then compared medaka peptides to silverside contigs using tblastn with soft masking and an e-value < 10-4. 19,230 of the 20,998 contigs had a reciprocal best hit to the medaka genome. Of these, 17,724 contigs mapped to one of the 24 medaka chromosomes.

Locations:
Five locations along the east coast of North America. Samples originally collected for the analysis were presented in Hice et al. (2012).

USA:Jekyll Island,Georgia 31.02 -81.43
USA:Oregon Inlet,NorthCarolina 35.77 -75.52
USA:Minas Basin,Gulf of Maine 45.2 -64.38
USA:Patchogue,NewYork 40.75 -73.00
Canada:Magdalen Island, Gulf of St. Lawrence 47.4 -61.85

[table of contents | back to top]

Related Publications

Hice, L. A., Duffy, T. A., Munch, S. B., & Conover, D. O. (2012). Spatial scale and divergent patterns of variation in adapted traits in the ocean. Ecology Letters, 15(6), 568–575. doi:10.1111/j.1461-0248.2012.01769.x
Methods

Therkildsen, N. O., & Palumbi, S. R. (2017). Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Molecular Ecology Resources, 17(2), 194–208. doi:10.1111/1755-0998.12593
Methods

Therkildsen, N. O., Wilder, A. P., Conover, D. O., Munch, S. B., Baumann, H., & Palumbi, S. R. (2019). Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. Science, 365(6452), 487–490. doi:10.1126/science.aaw7271
Methods

Wilder, A. P., Palumbi, S. R., Conover, D. O., & Therkildsen, N. O. (2020). Footprints of local adaptation span hundreds of linked genes in the Atlantic silverside genome. Evolution Letters, 4(5), 430–443. doi:10.1002/evl3.189
Results

[table of contents | back to top]

Related Datasets

IsRelatedTo

Therkildsen, N. O., Baumann, H. (2024) **Sample and genetic accession information for RNA-seq data from whole Atlantic silverside (Menidia menidia) larvae from two populations and their F1 hybrids reared under different temperatures in 2017.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2021-06-29 doi:10.26008/1912/bco-dmo.854887.1 [view at BCO-DMO]
Relationship Description: The RNA-seq dataset compares gene expression differences between two of the populations that we provide population genomic data for in the "Allele frequencies and FST estimates" dataset

Therkildsen, N. O., Baumann, H. (2024) **Sample information and genetic accession information for raw low-coverage genomic sequence reads from 248 different Atlantic silverside (Menidia menidia) collected along the east coast of North America between 2005 to 2007.** Biological and Chemical Oceanography Data Management Office (BCO-DMO). (Version 1) Version Date 2021-06-30 doi:10.26008/1912/bco-dmo.854878.1 [view at BCO-DMO]

Relationship Description: The raw low-coverage whole genome sequencing reads are the raw data used to generate the allele frequencies and FST estimates (for transcriptome regions only).

Wilder, A., Palumbi, S., Conover, D., & Overgaard Therkildsen, N. (2020). *Footprints of local adaptation span hundreds of linked genes in the Atlantic silverside genome* (Version 2) [Data set]. Dryad. https://doi.org/10.5061/DRYAD.JM63XSJ7W https://doi.org/10.5061/dryad.jm63xsj7w

[table of contents | back to top]

Parameters

Parameters for this dataset have not yet been identified

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina HiSeq 2000
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	Illumina HiSeq 2000 with paired-end 125 bp (for samples from GA, NY, GoM, and GSL)
Generic Instrument Description	General term for a laboratory instrument used for deciphering the order of bases in a strand of DNA. Sanger sequencers detect fluorescence from different dyes that are used to identify the A, C, G, and T extension reactions. Contemporary or Pyrosequencer methods are based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step.

Dataset-specific Instrument Name	Illumina NextSeq500
Generic Instrument Name	Automated DNA Sequencer
Dataset-specific Description	Illumina NextSeq500 paired-end 75 bp (for samples from NC).
Generic Instrument Description	General term for a laboratory instrument used for deciphering the order of bases in a strand of DNA. Sanger sequencers detect fluorescence from different dyes that are used to identify the A, C, G, and T extension reactions. Contemporary or Pyrosequencer methods are based on detecting the activity of DNA polymerase (a DNA synthesizing enzyme) with another chemoluminescent enzyme. Essentially, the method allows sequencing of a single strand of DNA by synthesizing the complementary strand along it, one base pair at a time, and detecting which base was actually added at each step.

[table of contents | back to top]

Project Information

Collaborative research: The genomic underpinnings of local adaptation despite gene flow along a coastal environmental cline (GenomAdapt)

Website: https://befel.marinesciences.uconn.edu/2018/03/07/research-news-new-nsf-grant-to-study-silverside-genes/

Coverage: Eastern coastline of North America

NSF Abstract:

Oceans are large, open habitats, and it was previously believed that their lack of obvious barriers to dispersal would result in extensive mixing, preventing organisms from adapting genetically to particular habitats. It has recently become clear, however, that many marine species are subdivided into multiple populations that have evolved to thrive best under contrasting local environmental conditions. Nevertheless, we still know very little about the genomic mechanisms that enable divergent adaptations in the face of ongoing intermixing. This project focuses on the Atlantic silverside (Menidia menidia), a small estuarine fish that exhibits a remarkable degree of local adaptation in growth rates and a suite of other traits tightly associated with a climatic gradient across latitudes. Decades of prior lab and field studies have made Atlantic silverside one of the marine species for which we have the best understanding of evolutionary tradeoffs among traits and drivers of selection causing adaptive divergence. Yet, the underlying genomic basis is so far completely unknown. The investigators will integrate whole genome sequencing data from wild fish sampled across the distribution range with breeding experiments in the laboratory to decipher these genomic underpinnings. This will provide one of the most comprehensive assessments of the genomic basis for local adaptation in the oceans to date, thereby generating insights that are urgently needed for better predictions about how species can respond to rapid environmental change. The project will provide interdisciplinary training for a postdoc as well as two graduate and several undergraduate students from underrepresented minorities. The findings will also be leveraged to develop engaging teaching and outreach materials (e.g. a video documentary and popular science articles) to promote a better understanding of ecology, evolution, and local adaptation among science students and the general public.

The goal of the project is to characterize the genomic basis and architecture underlying local adaptation in M. menidia and examine how the adaptive divergence is shaped by varying levels of gene flow and maintained over ecological time scales. The project is organized into four interconnected components. Part 1 examines fine-scale spatial patterns of genomic differentiation along the adaptive cline to a) characterize the connectivity landscape, b) identify genomic regions under divergent selection, and c) deduce potential drivers and targets of selection by examining how allele frequencies vary in relation to environmental factors and biogeographic features. Part 2 maps key locally adapted traits to the genome to dissect their underlying genomic basis. Part 3 integrates patterns of variation in the wild (part 1) and the mapping of traits under controlled conditions (part 2) to a) examine how genomic architectures underlying local adaptation vary across gene flow regimes and b) elucidating the potential role of chromosomal rearrangements and other tight linkage among adaptive alleles in facilitating adaptation. Finally, part 4 examines dispersal - selection dynamics over seasonal time scales to a) infer how selection against migrants and their offspring maintains local adaptation despite homogenizing connectivity and b) validate candidate loci for local adaptation. Varying levels of gene flow across the species range create a natural experiment for testing general predictions about the genomic mechanisms that enable adaptive divergence in the face of gene flow. The findings will therefore have broad implications and will significantly advance our understanding of the role genomic architecture plays in modifying the gene flow - selection balance within coastal environments.

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-1756316
NSF Division of Ocean Sciences (NSF OCE)	OCE-1756751

[table of contents | back to top]