Sample collection and sequence accession information for Zostera marina whole genome resequencing from specimens collected at 16 geographic locations worldwide in 2017

Website: https://www.bco-dmo.org/dataset/924852
Data Type: Other Field Results
Version: 1
Version Date: 2024-04-10

Project
» Using genomics to link traits to ecosystem function in the eelgrass Zostera marina (ZosteraEcoGenomics)

Contributors	Affiliation	Role
Stachowicz, John J.	University of California-Davis (UC Davis)	Principal Investigator
Rauch, Shannon	Woods Hole Oceanographic Institution (WHOI BCO-DMO)	BCO-DMO Data Manager

Abstract
This dataset includes sample collection and sequence accession information for Zostera marina whole genome resequencing from specimens collected at 16 geographic locations worldwide in 2017. Sequence accessions are housed in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA).

Table of Contents

Coverage

Location: Worldwide at 16 sites within shallow seagrass beds (< 2 meters deep)
Spatial Extent: **N**:67.267997 **E**:7.493273 **S**:32.713756 **W**:-117.225474
Temporal Extent: 2017-08-15 - 2017-10-30

Methods & Sampling

We conducted a range-wide sampling collection of 190 *Zostera marina* specimens from 16 geographic locations (Yu et al. 2023 Nature Plants). The chosen populations feature a mix of sexual and vegetative reproduction with the exception of mostly vegetative reproduction at the sites PO and NN, apparent through extended clones. Chosen locations were a subset of the Zostera Experimental Network sites that were previously analysed using 24 microsatellite loci. Although a sampling distance of >2 meters was maintained to reduce the likelihood of collecting the same genet/clone twice, this was not always successful and thus provided an estimate of local clonal diversity. Plant tissue was selected from the basal meristematic part of the shoot after peeling away the leaf sheath to minimize epiphytes (bacteria and diatoms), frozen in liquid nitrogen and stored at -80 degrees Celsius (°C) until DNA extraction. Quality control was performed following Joint Genome Institute guidelines (https://jgi.doe.gov/wp-content/uploads/2013/11/Genomic-DNA-Sample-QC.pdf). Plate-based DNA library preparation for Illumina sequencing was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using Kapa Biosystems library preparation kit. About 200 nanograms (ng) of sample DNA was sheared to a length of around 600 base pairs (bp) using a Covaris LE220 focused

ultrasonicator. Selected fragments were end-repaired, A-tailed, and ligated with sequencing adaptors containing a unique molecular index barcode. Libraries were quantified using KAPA Biosystems' next-generation sequencing library qPCR-kit on a Roche LightCycler 480 real-time PCR instrument. Quantified libraries were then pooled together and prepared for sequencing on the Illumina HiSeq2500 sequencer using TruSeq SBS sequencing kits (v4) following a 2 × 150 bp indexed run recipe to a targeted depth of approximately 40x coverage. The quality of the raw reads was assessed by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and visualized by MultiQC58. BBDuk (https://jgi.doe.gov/data-and-tools/software-tools/bbtools/bb-tools-user-guide/bbduk-guide/) was used to remove adapters and for quality filtering, discarding sequence reads (1) with more than one 'N' (maxns = 1), (2) shorter than 50 bp after trimming (minlength = 50), and (3) with average quality <10 after trimming (maq = 10). FastQC and MultiQC were used for a second round of quality check for the clean reads. Sequencing coverage and mapping rate were calculated for each sample (see Yu et al. 2023 for details).

Data Processing Description

To analyse genetic loci present throughout the global distribution range of eelgrass, we focused on identifying core genes that are present in genomes of all individuals. To do so, each of the 190 ramets were de novo assembled using HipMer (k = 51). To categorize, extract and compare core and variable (shell and cloud) genes, primary transcript sequences (21,483 gene models) from the *Z. marina* reference (V3.1; ref. 19) were aligned using BLAT using default parameters to each de novo assembly. Genes were considered present if the transcript aligned with either (1) >60% identity and >60% coverage from a single alignment or (2) >85% identity and >85% coverage split across three or fewer scaffolds. Individual presence–absence-variation calls were combined into a matrix to classify genes into core, cloud and shell categories based on their observation across the population.

The total number of genes considered was 20,100. Because identical genotypes and fragmented, low-quality assemblies can bias and skew presence–absence-variation analyses, only 141 single representatives of clones and ramets with greater than 17,500 genes were kept to ensure that only unique, high-quality assemblies were retained. Genes were classified using discriminant analysis of principal components into cloud, shell and core gene clusters based on their frequency. Core genes were the largest category, with 18,717 genes that were on average observed in 97% of ramets.

SNP mapping, calling and filtering The quality-filtered reads were mapped against the chromosome-level *Z. marina* reference genome V3.1 using BWA MEM62. The alignments were converted to BAM format and sorted using Samtools. The Mark- Duplicates module in GATK4 was used to identify and tag duplicate reads in the BAM files. The mapping rate for each genotype was calculated using Samtools (Supplementary Data Table 2 of Yu et al. 2023). HaplotypeCaller (GATK4) was used to generate a Genomic Variant Call Format (GVCF) file for each sample, and all the GVCF files were combined by CombineGVCFs (GATK4). GenotypeGVCFs (GATK4) was used to call genetic variants. BCFtools64 was used to remove SNPs within 20 base pairs of an indel or other variant type, as these variant types may cause erroneous SNPs calls. VariantsToTable (GATK4) was used to extract INFO annotations. SNPs meeting one or more than one of the following criteria were marked by VariantFiltration (GATK4): MQ < 40.0; FS > 60.0; QD < 10.0; MQRandSum > 2.5 or MQRandSum < −2.5; Read- PosRandSum < −2.5; ReadPosRandSum > 2.5; SOR > 3.0; DP > 10,804.0 (2 × average DP). Those SNPs were excluded by SelectVariants (GATK4). A total of 3,975,407 SNPs were retained. VCFtools65 was used to convert individual genotypes to missing data when GQ < 30 or DP < 10. Individual homozygous reference calls with one or more reads supporting the variant allele, and individual homozygous variant calls with ≥1 read supporting the reference, were set as missing data. Only bi-allelic SNPs were kept (3,892,668 SNPs). To avoid the reference-genome-related biases, due to the large Pacific–Atlantic genomic divergence, we focused on the 18,717 core genes that were on average observed in 97% of ramets. Bedtools66 was used to find overlap between the SNPs and the core genes, and only those SNPs were kept (ZM_HQ_SNPs, 763,580 SNPs). Genotypes that were outside our custom quality criteria were represented as missing data.

Possible parent-descendant pairs under selfing as well as clonemates were detected based on the shared heterozygosity. To ensure that all genotypes assessed originated by random mating, ten ramets showing evidence for selfing were excluded. Seventeen multiple sampled clonemates were also excluded. After the exclusion of 37 samples owing to missing data, selfing, or clonality, 153 samples were left for further analyses.

The chloroplast genome was de novo assembled by NOVOPlasty. The chloroplast genome of *Z. marina* was represented by a circular molecule of 143,968 bp with a classic quadripartite structure: two identical inverted repeats (IRa and IRb) of 24,127 bp each, a large single-copy region of 83,312 bp, and a small single-copy region of 12,402 bp. All regions were equally taken into SNP calling analysis except for 9,818 bp encoding 23S

and 16S ribosomal RNAs due to bacterial contamination in some samples. The raw Illumina reads of each individual were aligned by BWA MEM to the assembled chloroplast genome. The alignments were converted to BAM format and then sorted using Samtools. Genomic sites were called as variable positions when the frequency of variant reads was >50% and the total coverage of the position was >30% of the median coverage (174 variable positions). Then 11 positions likely related to microsatellites and 12 positions reflecting minute inversions caused by hairpin structures were removed from the final set of variable positions for the haplotype reconstruction (151 SNPs). For the phylogenetic tree reconstruction, we further selected 108 SNPs that represent parsimony-informative sites (that is, no singletons).

BCO-DMO Processing Description

- Created a lookup table to match SRA accession number to BioProject by running script: https://gist.github.com/adyork/07ae4060e3017ff8cddd141dbe2ffe3a#file-sra...
- Saved this lookup table as "srp_and_bioproject.csv".
- Imported original file "Stachowicz_AccessionNumbers.csv" into the BCO-DMO system.
- Imported file "srp_and_bioproject.csv".
- Joined the two files together, adding the BioProject numbers to the primary data table, matching on the SRA accession number.
- Renamed fields to comply with BCO-DMO naming conventions.
- Saved the final file as "924852_v1_z_marina_whole_genome_resequencing.csv".

[table of contents | back to top]

Data Files

File	
924852_v1_z_marina_whole_genome_resequencing.csv	(Comma Separated Values (.csv), 43.80 KB) MD5:70c9335afc5862ee5e691ce3bfd840eb
Primary data file for dataset ID 924852, version 1	

[table of contents | back to top]

Related Publications

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc
Software

Bushnell, B. (2014). BBTools software package. http://bbtools.jgi.doe.gov
Software

Yu, L., Khachaturyan, M., Matschiner, M., Healey, A., Bauer, D., Cameron, B., Cusson, M., Emmett Duffy, J., Joel Fodrie, F., Gill, D., Grimwood, J., Hori, M., Hovel, K., Hughes, A. R., Jahnke, M., Jenkins, J., Keymanesh, K., Kruschel, C., Mamidi, S., … Reusch, T. B. H. (2023). Ocean current patterns drive the worldwide colonization of eelgrass (Zostera marina). Nature Plants, 9(8), 1207–1220. https://doi.org/10.1038/s41477-023-01464-3
Results

[table of contents | back to top]

Parameters

Parameter	Description	Units
SampleID	Sample identifier	unitless
BioProject	NCBI BioProject number	unitless
Library_Code	Library Code	unitless
Sample_Description	Sample Description	unitless
Location_Name	Location name	unitless
Lat	Site latitude	decimal degrees
Long	Site longitude; negative values = West	decimal degrees
SRA_accession	NCBI SRA Study Accession identifier	unitless
Library_type	Library type	unitless
Sampled_tissue	Description of sampled tissue	unitless
Instrument	Type of sequencing instrument	unitless
reads	read length	unitless
Number_of_bases_Raw_fastq	Total number of base pairs sequenced in raw files	count
Coverage_Raw_fastq	Genome Coverage (times covered) - Raw	genomes
Number_of_bases_Clean_fastq	Genome Coverage (times covered) - Cleaned	count
Coverage_Clean_fastq	Fraction of the genome covered by cleaned sequencing	genomes
Mapped_pcnt	Percent of reads mapped to the genome	percent
Properly_paired_pcnt	Percent properly paired	percent

[table of contents | back to top]

Instruments

Dataset-specific Instrument Name	Illumina HiSeq2500 sequencer
Generic Instrument Name	Automated DNA Sequencer
Generic Instrument Description	A DNA sequencer is an instrument that determines the order of deoxynucleotides in deoxyribonucleic acid sequences.

[table of contents | back to top]

Project Information

Using genomics to link traits to ecosystem function in the eelgrass Zostera marina (ZosteraEcoGenomics)

Coverage: In Zostera marina beds worldwide, including western and eastern margins of both the Atlantic and Pacific Oceans. Project centered in Bodega Bay, CA 38.31 N; 123.059 W

NSF Award Abstract:
Seagrass ecosystems provide important services to coastal regions, including primary production, carbon storage, nutrient cycling, habitat for fisheries species, and erosion control. At the same time, eelgrass is threatened by direct destruction, pollution, and other human impacts on the environment. We know that genetic diversity in eelgrass enhances seagrass bed growth and persistence, but application of this knowledge to restoration and conservation is limited. This work will guide restoration programs by considering what specific aspects of diversity are important to conservation and restoration of seagrass ecosystems, helping to guide the selection of source material to improve restoration success (which is often low). The project integrates the effects of multiple components of diversity and clarifies the extent to which genetic and ecological uniqueness can predict ecosystem functions.

Intellectual Merit: Genetic diversity as measured by the number of genetically distinct individuals (genets) in an assemblage influences critical ecosystem functions in a wide range of ecosystems. Functional diversity, the presence of key traits, or population flexibility to respond to environmental change are all potential mechanisms underlying these patterns, but distinguishing among them requires a clear link between genetic diversity and the phenotypes present in an assemblage. The investigators, and others, have previously demonstrated that genet diversity in eelgrass (Zostera marina) increases stand productivity, animal community diversity, and resilience to environmental change. These genet diversity effects are associated with increases in genetically determined trait diversity. Predicting trait diversity without having to measure traits of every genet remains a major barrier to wider application of functional diversity approaches in restoration and management. In this project, the investigators assess the association between Single Nucleotide Polymorphisms (SNPs) across the genome and performance-related traits that we will measure at the individual, population, and seascape-scale. They also assess environmental correlates of trait differentiation from field sampling. Finally, the research team will compare the predictive power of genomic SNP diversity versus other metrics of intraspecific diversity for the functioning (productivity, invertebrate abundance) of field planted eelgrass assemblages. If genomic variation can reliably be used to predict functional traits, then the value of genomic sequencing efforts for informing management will be greatly enhanced. Broader Impacts: Seagrass restoration and mitigation is currently of major interest in California and elsewhere and the project results will inform current initiatives regarding eelgrass management in California through the state's Ocean Protection Council. In addition to recruiting individual students from diverse backgrounds to work on the project, the project broadens participation of students in STEM fields through its partnership with three existing outreach/training programs at UC Davis.

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-1829976

[table of contents | back to top]