Supplementary Materials

Fine-scale spatial and temporal genomic variation among Dungeness crab *Cancer magister* larval recruits in the California Current Ecosystem

Authors: Elizabeth M. J. Lee*, Kathleen G. O'Malley

State Fisheries Genomics Lab, Coastal Oregon Marine Experiment Station, Department of Fisheries and Wildlife, Hatfield Marine Science Center, Oregon State University, Newport, Oregon 97365, USA

*Corresponding author email: elizabethmjlee@gmail.com

Text S1. Bioinformatic Methods

S1.1 Filtering and Demultiplexing Sequences

The raw sequence reads from the Dungeness crab megalopae genotyping-by-sequencing (GBS) libraries were assessed and filtered for quality before downstream genetic analyses were conducted. First, the program 'fastqc' v. 0.11.3, (Andrews 2010) was used to assess the quality of the Illumina raw reads. Next, the program 'cutadapt' v. 1.5 (Martin 2011) was used to trim any Illumina primers from the reads. Then, the program 'kraken' v. 2.0.6-beta (Wood & Salzberg 2014) was used to identify and remove any reads that matched bacterial sequences in the standard 'kraken' database (NCBI RefSeq; O'Leary et al. 2015). Finally, the 'stacks' v. 2.2 (Catchen et al. 2011, Catchen et al. 2013) program 'process_radtags' was implemented to demultiplex the paired-end (-P) reads for each sample based on the unique barcodes (-inline_null) while filtering for quality. In filtering for quality with 'process_radtags', sequences with an uncalled base were removed (-c), barcodes and restriction enzyme sites with one mismatch difference from a true barcode or a restriction enzyme site sequence were corrected (-r), and sequences with an average quality score below 20 within a sliding window of 15% of the read length were removed (-s 20, -w 0.15, -q).

S1.2 Assembling and Filtering Loci

Since there was no reference genome for the Dungeness crab at the time of the study, the 'stacks' v. 2.2 (Catchen et al. 2011, Catchen et al. 2013) 'denovo_map.pl' pipeline was used to de novo assemble the filtered reads into putative loci (Rochette & Catchen 2017). The 'ustacks' program (-M 2, -m 4, -N 4) was first executed to identify putative loci within each sample. The 'cstacks' program was then executed to compare putative loci across all 2017 and 2018 megalopae and compile a catalogue of consensus loci. The 'sstacks' program was used to match

the loci of each individual to all loci within the compiled catalogue of loci. Finally, the 'tsv2bam' program was used to transpose the loci data and the 'gstacks' program was used to assemble contigs and call single nucleotide polymorphisms (SNPs) within each locus.

S1.3 Identifying and Calling SNPs

To test for intra- and inter- annual genetic differentiation across both sampling sites in 2017 and 2018, the 'stacks' program 'populations' was used to produce a curated set of high quality loci across all collection timepoints. Loci were retained if they were present within >70% of individuals (-r 0.7) from each collection timepoint (n=8): both sites, both years, and both collection timepoints within each year (-p8). The SNPs retained had minor allele frequencies >0.05 (--min_maf 0.05) and only the first SNP of each locus was selected for use in downstream analyses to maintain independence (--write_single_snp). The program 'plink' v. 1.07 (Purcell et al. 2007) was used to determine if any loci were in linkage disequilibrium (LD; $r^2 > 0.80$). The loci with a proportion of heterozygotes greater than 0.6 or an allele depth ratio deviation greater than ±5.0 were identified as putative paralogous sequence variants (PSVs; McKinney et al. 2017). Using the program 'vcftools' v. 0.1.13 (Danecek et al. 2011), loci with read depths less than 10 were identified and removed, and further loci were removed if they had a read depth mean or read depth standard deviation (SD) greater than 1.5 times the interquartile range above the upper quartile (i.e. outlier read depths). One locus from each pair of loci in LD, all loci identified as putative PSVs, and all loci with low or high read depths were compiled into a blacklist, and the 'stacks' program 'populations' was rerun to exclude the blacklisted loci. Individuals missing >25% of data across all loci were removed. The final set of polymorphic loci with corresponding single SNP genotypes were compiled with the 'populations' program and output as a variant call format (VCF) file.

Text S2. Bioinformatic Results

S2.1 Filtering and Demultiplexing Sequences

Sequencing the four GBS libraries containing 378 megalopae resulted in a mean of 375,329,320 (± 16,841,901 SD) read pairs per library. A mean of 240,520,948 (± 30,997,035 SD) read pairs remained after removing sequences that contained Illumina adapters (library inserts less than 150 bp) and a mean of 227,282,896 (± 28,137,115 SD) read pairs remained after removing reads containing bacterial sequences. After demultiplexing and further filtering the raw reads with 'stacks' program 'process_radtags', a mean of 3,310,501 (± 921,916 SD) reads per megalopa were retained (Figure S1). Four individuals were removed due to poor sequencing. When the reads were assembled with 'stacks', a catalogue of 259,928 loci was formed. After filtering, 1,391 loci remained for analysis (Table S2). Three individuals were removed from the dataset due to >25% missing data. Overall, the dataset had a mean estimated heterozygote miscall rate of <10%.

Literature Cited

Andrews S (2010) FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc (accessed 1 Dec 2018)

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. G3 1:171-182

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124-3140

Danecek P, Auton A, Abecasis G, Albers CA and others (2011) The variant call format and VCFtools. Bioinformatics 27:2156-2158

Flanagan SP, Jones AG (2017) Constraints on the FST–hetero zygosity outlier approach. J Hered 108:561-573

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180:977-993

Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10-12

McKinney GJ, Waples RK, Seeb LW, Seeb JE (2017) Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping□by□sequencing data from natural populations. Mol Ecol Resour 17:656-669

Milligan BG (2003) Maximum-likelihood estimation of relatedness. Genetics 163: 1153-1167

O'Leary NA, Wright MW, Brister JR, Ciufo S and others (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44:733-745

Purcell S, Neale B, Todd-Brown K, Thomas L and others (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genets 81:559-575

Rochette NC, Catchen JM (2017) Deriving genotypes from RAD-seq short-read data using Stacks. Nat Protoc 12:2640

Wang J (2011) Coancestry: a program for simulating, estimating and analysing relatedness and inbreeding coefficients. Mol Ecol Resour 11:141-145

Whitlock MC, Lotterhos KE (2015) Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST. Am Nat 186(Suppl 1): S24-S36

Wood DE, Salzberg SL (2014) Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:46

Figures

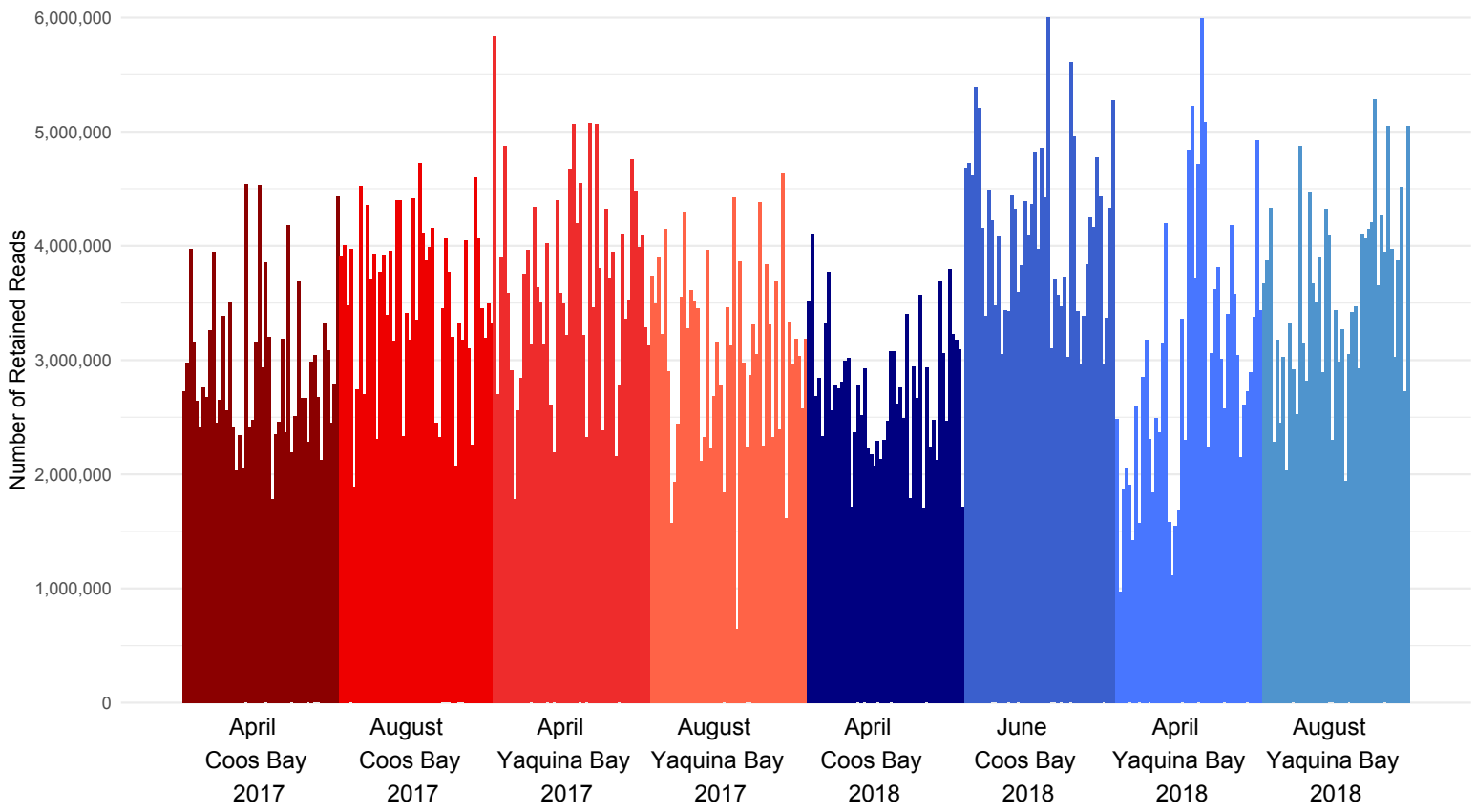

Figure S1. The number of sequence reads retained for each individual megalopa after demultiplexing and filtering sequence reads through 'stacks' v. 2.2 (Catchen et al. 2011, Catchen et al. 2013) 'process_radtags' program. Collection timepoints are separated by color.

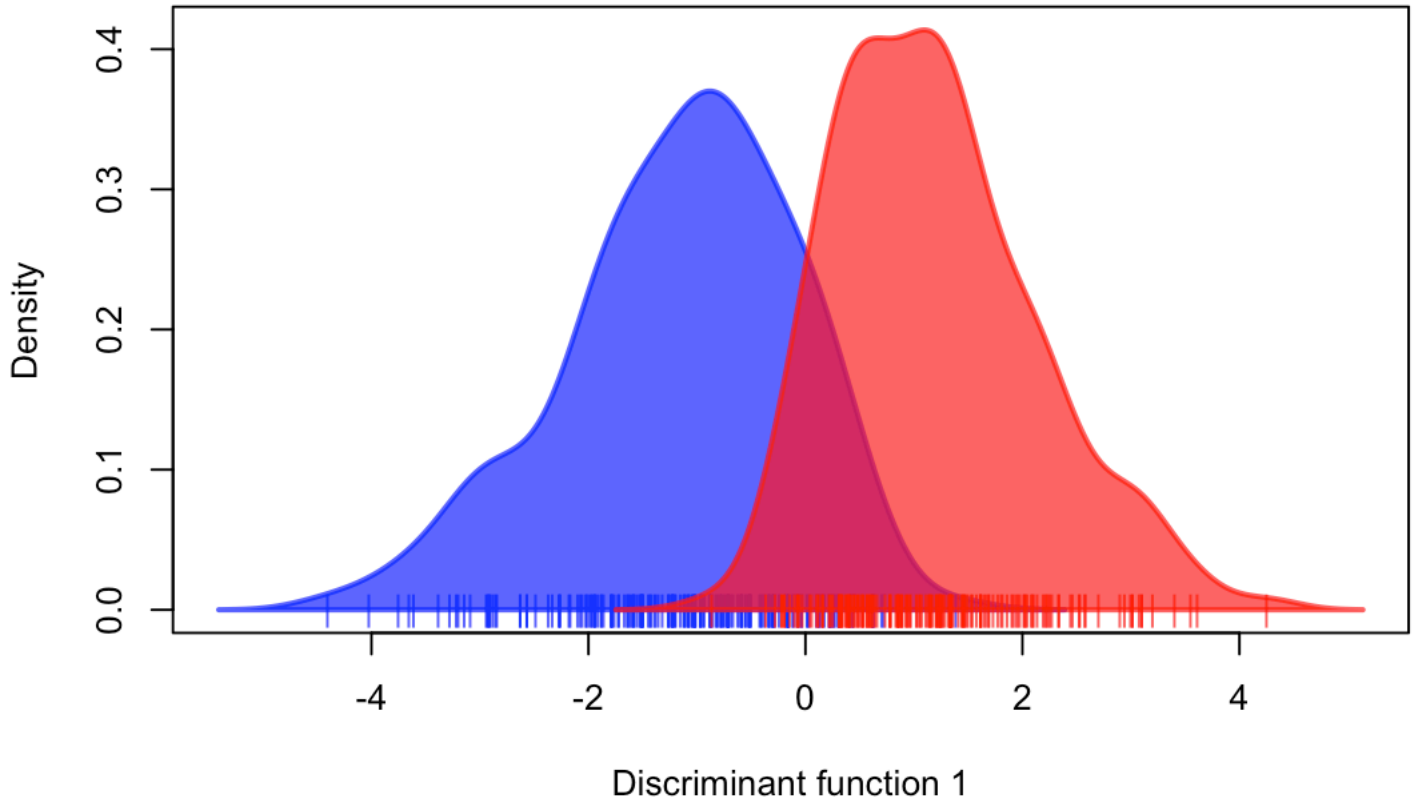

Figure S2. Discriminate analysis of principal components (DAPC) for expected-season (blue) and late-season (red) megalopae recruits in 2017 and 2018 in Coos Bay, Oregon using both presumably neutral (1,389) and putatively adaptive (2) loci for principal components 1 through 10.

Tables

Table S1. Reported Pearson's correlation coefficient for each relatedness estimator tested with simulated Dungeness crab megalopae dyads and implemented in 'coancestry' v. 1.0.1.9 (Wang 2011) (1,000 dyads per 100 datasets). The Milligan (2003) relatedness estimator was chosen for the study.

Relatedness Estimator	Pearson's Correlation Coefficient
Li et al. (1993)	0.960
Lynch and Ritland (1999)	0.959
Milligan (2003)	0.972
Queller and Goodnight (1989)	0.963
Ritland (1996)	0.945
Wang (2002)	0.958
Wang (2007)	0.971

Table S2. The number of loci removed at each filtering step.

Filtering Step	Number of Loci
Original number of loci identified with 'stacks'	**1,767**
Removed based on linkage disequilibrium (LD)	0
Removed based on presence of Paralogous Sequence Variant (PSV)	340
Removed based on low read depth	11
Removed based on high read depth	25
Final number of retained loci**	**1,391**

***Mean read depth = 171*

Table S3. Reported number of the 1,391 loci identified through the STACKS assembly of sequence reads which deviated from Hardy-Weinberg proportions (HWP) within each collection time point. The number of loci with observed heterozygote (Het) deficiency or excess are reported.

Year	Site	Month	Recruitment Season	n	HWP Deviation	# Het Deficiency	# Het Excess
2017	Coos Bay	April	Expected	47	57	57	0
		August	Late	47	63	63	0
	Yaquina Bay	April	Expected	48	82	82	0
		August	Late	48	90	89	1
2018	Coos Bay	April	Expected	47	58	58	0
		June	Expected	44	60	60	0
	Yaquina Bay	April	Expected	45	54	54	0
		August	Late	45	69	69	0

Table S4. Number of putatively adaptive loci identified with three outlier detection programs: 'bayescan' v. 2.1 (Foll & Gaggiotti), 'outflank' v. 0.2 (Whitlock & Lotterhos 2015), and 'fsthet' v. 1.0.1 (Flanagan & Jones 2017).

	Number of Loci
Total Loci	**1,391**
Loci identified by 'bayescan'	2
Loci identified by 'outflank'	21
Loci identified by 'fsthet'	41
Final putatively adaptive Loci**	**2**
Final Neutral Loci	**1,389**

***Loci identified by all three outlier detection programs*

Table S5. Consensus sequences of the two putatively adaptive loci identified by all three outlier detection programs ('bayescan', 'outflank', and 'fsthet').

Outlier Loci	Consensus Sequence
CLocus_24	TGCAGGAAAAAAGAATAAAAAAAATTTTTTTTTTCATGAGTCTTCATCCAATCCTTCTTCCAAACC ATTAATATCTGCTTGTGTTGCATGGTATTCTGTGCGCTGAAAATTGCTGTAATATACTAAGACTTGA ACCGCCGCTGGTACGTACTGTTGAAGGCNNNNNNNNTGAGGTTGGTAGTAGCCGTCAGCACCGTTC ATGTAGTCGGAATACTCGTAGGCGTGCGGCGGGTAGGACATGTTTGGATAGAAGCTCGGGTACCCT TGCTGCGTTGGGGAGAAGTAGTGAAGCATCTTGGAAATGCCATTGGGGACATCCGGAA
CLocus_3063	TGCAGGACTTTCGCGGCCACACCTGAAAGCAGCACGCAACACACTGGTCCAAAATATCCCTACCTT TCGACCTTCAACTTTATCTTTCCTCAACCTCGATTCTTGGTAGTGGAAGGTCATCTAGCCTGGGAAT AAAGGATTGTTTNNNATTGTCAGTTTTCTTTGACATTGGTTTACGGCAAGGCTCTCAACTCTTATAA GGTGTTCGTCGACGTCTATGCTTTTGGAGTACCCG

Table S6. Pairwise F_{ST} estimates based on variation at the 1,391 neutral and putatively adaptive loci identified within Dungeness crab megalopae collected in 2017 and 2018. No comparisons were significant after false discovery rate (FDR) correction. Late-season collection timepoints (i.e. August) are shaded in gray. Pairwise F_{ST} comparisons between expected- and late-season samples are in boldface. Intra-annual F_{ST} estimates are shaded blue while inter-annual F_{ST} estimates are not shaded.

			2017				**2018**			
			Coos Bay		**Yaquina Bay**		**Coos Bay**		**Yaquina Bay**	
			April	**August**	**April**	**August**	**April**	**June**	**April**	**August**
2017	**Coos Bay**	**April**								
		August	**0.0021**							
	Yaquina Bay	**April**	0.0004	**0.0017**						
		August	**0.0021**	0.0004	**0.0018**					
2018	**Coos Bay**	**April**	0.0000	**0.0018**	0.0006	**0.0017**				
		June	0.0008	**0.0019**	0.0000	**0.0015**	0.0000			
	Yaquina Bay	**April**	0.0000	**0.0022**	0.0004	**0.0021**	0.0000	0.0007		
		August	**0.0000**	0.0014	**0.0009**	0.0004	**0.0010**	**0.0000**	**0.0000**	