A chromosome-level reference genome for the giant pink sea star, *Pisaster brevispinus*, a species severely impacted by wasting

Melissa B. DeBiasse[1], Lauren M. Schiebelhut[1], Merly Escalona[2], Eric Beraut[3], Colin Fairbairn[3], Mohan P.A. Marimuthu[4], Oanh Nguyen[4], Ruta Sahasrabudhe[4], Michael N Dawson[1]

[1]Department of Life and Environmental Sciences, University of California, Merced, CA 95343, USA

[2]Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA

[3]Ecology & Evolutionary Biology Department, 1156 High St, University of California Santa Cruz, Santa Cruz, CA 95064, USA

[4]DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California Davis, CA 95616, USA

Corresponding author email

Melissa DeBiasse <melissa.debiasse@gmail.com>

Abstract

Efforts to protect the ecologically and economically significant California Current Ecosystem from global change will greatly benefit from data about patterns of local adaptation and population connectivity. To facilitate that work, we present a reference quality genome for the giant pink sea star, *Pisaster brevispinus*, a species of ecological importance along the Pacific west coast of North America that has been heavily impacted by environmental change and disease. We used Pacific Biosciences HiFi long sequencing reads, and Dovetail Omni-C proximity reads to generate a highly contiguous genome assembly of 550Mb in length. The assembly contains 127 scaffolds with a contig N50 of 4.6Mb and a scaffold N50 of 21.4Mb; the BUSCO completeness score is 98.70%. The *P. brevispinus* genome assembly is comparable to the genome of the sister species *P. ochraceus* in size and completeness. Both *Pisaster* assemblies are consistent with previously published karyotyping results showing sea star genomes are organized into 22 autosomes. The reference genome for *P. brevispinus* is an important first step toward the goal of producing a comprehensive, population genomics view of ecological and evolutionary processes along the California coast. This resource will help scientists, managers, and policy makers in their task of understanding and protecting critical coastal regions from the impacts of global change.

Keywords

Asteroidea, California Conservation Genomics Project (CCGP), Echinodermata, global change, marine invertebrate, California Current Ecosystem (CCE)

Introduction

The California Current Ecosystem (CCE) is a dynamic and complex region of high ecological and economic value (Weber, et al. 2021). A key component of protecting the value of the CCE from the negative impacts of global change is a comprehensive understanding of the connections and interactions of the species that exist here. Decades of population biology and ecology research has been conducted in the Pacific Northwest generally (Menge, et al. 2019), and in California specifically (Connell 1972; Sagarin, et al. 1999; Blanchette, et al. 2008; Sanford, et al. 2019). In recent years, studies have revealed populations of up to two dozen species negatively impacted, in some cases being locally extirpated, by environmental stressors including increasing temperature change, harmful algal blooms, and disease outbreaks (Jurgens, et al. 2015; Harvell and Lamb 2020). Species loss and the subsequent breakdown of important interactions are detrimental to CCE function and could ultimately include region-wide ecosystem collapse (Burt, et al. 2018; McPherson, et al. 2021).

Addressing the intensifying threats to coastal ecosystems requires collaborative, interdisciplinary efforts by all stakeholders. A critical part of this process includes increasing genomic resources, which can be used to address ecological questions and inform conservation decisions, uniting scientists, managers, and policy makers, and complementing decades of foundational field work. Analyzing genomic data for coastal species will reveal how interspecific variation in sequence (e.g., nucleotide polymorphism) and structure (e.g., chromosomal inversions) relate to variation in susceptibility to environmental stress, furthering the goal of preserving natural resources in California and beyond (Formenti, et al. 2022; Shaffer, et al. 2022).

Sea stars (Echinodermata, Asteroidea) are among the taxa most severely impacted by ongoing environmental change (Montecino-Latorre, et al. 2016). Sea stars are significant members of intertidal and subtidal communities with some species acting as keystone species, a concept inspired by the role of the ochre sea star, *Pisaster ochraceus*, in the Northeast Pacific (Paine 1966; Schultz, et al. 2016). Of the twenty or more species impacted by the geographically and phylogenetically broad sea star wasting outbreak in 2013, *P. brevispinus*, a congener of *P. ochraceus*, was one of the most severely impacted, with widespread wasting and precipitous population declines (Montecino-Latorre, et al. 2016, Dawson et al. in review). Recent research has shown that losing sea stars from coastal ecosystems has cascading detrimental effects (Burt, et al. 2018), further supporting their importance to nearshore communities, and motivating efforts to conserve the biodiversity that remains.

Here we present the reference genome assembly for the giant pink sea star, *P. brevispinus* (Forcipulatida, Asteriidae) (Stimpson), a large-bodied, fast-moving sea star with 5 rays (i.e., arms) found in the low intertidal zone, but more commonly in the neritic zone on soft substrates from circa Ensenada, Baja California, Mexico, to Sitka, Alaska, USA (Figure 1, Morris, et al. 1980; Costello, et al. 2013; Beas- Luna, et al. 2020). They are gonochoristic broadcast spawners (Morris, et al. 1980) with an estimated larval duration of 76-266 days (Strathmann 1987). *Pisaster brevispinus* is an exceptional predator: it can extend tube feet on the oral disc into the sediment as far as the length of the sea star's radius (up to ~16cm), pulling clams and other prey to the surface for consumption (Morris, et al. 1980). The reference genome produced here will contribute to our understanding of ecological and evolutionary patterns through comparisons with other taxa and has the potential to

reveal hotspots of genetic diversity, connectivity, and species-associations that shape population dynamics and ecosystems along the California coast (Shaffer, et al. 2022).

Methods

Biological Materials

An adult *P. brevispinus*, 118mm radius (arm tip to disc center), was collected from a sandstone platform at 11-13m depth at Terrace Point, Santa Cruz, CA, USA (36.94487, -122.06429) on 13 October 2020 by Shannon Myers. The voucher specimen (M0D059179O) is archived in the Dawson Lab at the University of California, Merced, USA.

Nucleic acid extraction, library preparation, and sequencing

We extracted high molecular weight (HMW) genomic DNA (gDNA) from 28mg of tube foot tissue using Nanobind Tissue Big DNA kit (Pacific BioSciences - PacBio) following the manufacturer's instructions with the following minor modification: we centrifuged tissue homogenate at 18000g (instead of recommended 1500g) during the second wash because faster speeds were required to remove the excess wash buffer retained in the tube foot tissue during homogenization. We assessed DNA purity using absorbance ratios (260/280 = 1.87 and 260/230 = 2.47) on a NanoDrop ND-1000 spectrophotometer. We quantified DNA yield (210ng/µl; 20µg total) using the Quantus Fluorometer QuantiFluor ONE dsDNA Dye assay.

We constructed the PacBio HiFi library using the SMRTbell Express Template Prep Kit v2.0 according to the manufacturer's instructions. We sheared 13.1µg of HMW gDNA to an average size distribution of ~18Kb using Diagenode's Megaruptor 3 system. We quantified the sheared DNA using the Quantus Fluorometer and checked the size distribution using the Agilent Femto Pulse. We concentrated the sheared gDNA using 0.45X of AMPure PB beads followed by quantification using a Quantus Fluorometer. We used 6µg of sheared, concentrated DNA as input for the removal of single-strand overhangs at 37ºC for 15 minutes, followed by further enzymatic steps of DNA damage repair at 37ºC for 30 minutes, end repair and A-tailing at 20ºC for 10 minutes and 65ºC for 30 minutes, ligation of overhang adapter v3 at 20º for 1 hour and 65ºC for 10 minutes to inactivate the ligase, and nuclease treatment of SMRTbell library at 37º for 1 hour to remove damaged or non-intact SMRTbell templates. We purified and concentrated the SMRTbell library with 0.8X Ampure PB beads for size selection using the BluePippin system. We purified the input of 3.2µg purified SMRTbell library to load into the Blue Pippin 0.75% Agarose Cassette using cassette definition 0.75% DF Marer S1 3-10Kb Improved Recovery for the run protocol. We collected fragments greater than 7Kb from the cassette elution well and purified and concentrated the size-selected SMRTbell library with 0.8X AMPure beads.

We performed proximity ligation using the Dovetail™ Omni-C™ Kit according to the manufacturer's protocol with slight modifications. First, we thoroughly ground the specimen tissue with a mortar and pestle in liquid nitrogen. Subsequently, chromatin was fixed in place in the nucleus. We passed the

suspended chromatin solution through 100µm and 40µm cell strainers to remove large debris. We digested fixed chromatin under various conditions of DNase I until a suitable fragment length distribution of DNA molecules was obtained. We repaired and ligated the chromatin ends to a biotinylated bridge adapter followed by proximity ligation of adapter containing ends. After proximity ligation, crosslinks were reversed, and the DNA purified from proteins. We treated the purified DNA to remove biotin that was not internal to ligated fragments. We generated a library with an Illumina compatible y-adaptor using the NEB Ultra II DNA Library Prep kit and captured biotin-containing fragments using streptavidin beads. We split the post-capture product into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The 20.5 kb average HiFi SMRTbell library was sequenced using one 8M SMRT Cell, Sequel II sequencing chemistry 2.0, and 30-hour movies on a PacBio Sequel II sequencer. The Omni-C library was sequenced on an Illumina NovaSeq platform to generate approximately 100 million 2 x 150Bp read pairs per gigabase of genome size.

Nuclear and mitochondrial genome assemblies

We assembled the genome of the giant pink sea star following the California Conservation Genomics Project (CCGP) assembly protocol Version 4.0, introduced in (Lin, et al. 2022). The difference between versions relies on the output sequences from HiFiasm [Version 0.16.1-r375] (Cheng et al. 2020) that are used to generate the final assembly (See Table 1 for assembly pipeline and relevant software). The final output corresponds to a dual or partially phased diploid assembly (http://lh3.github.io/2021/10/10/introducing-dual-assembly).

We initially removed remnant adapter sequences from the PacBio HiFi dataset using HiFiAdapterFilt [Version 1.0] (Sim, et al. 2022) and generated the initial diploid assembly with the filtered PacBio and the Omni-C data using HiFiasm. We tagged output haplotype 1 as the primary assembly, and output haplotype 2 as the alternate assembly. Next, we identified sequences corresponding to haplotypic duplications on the primary assembly with purge_dups [Version 1.2.6] (Guan, et al. 2020) and transferred them to the alternate assembly. We scaffolded both assemblies using the Omni-C data with SALSA [Version 2.2] (Ghurye, et al. 2017; Ghurye, et al. 2019).

Both assemblies were manually curated by generating and analyzing Omni-C contact maps and breaking the assemblies when major misassemblies were found. No further joins were made after this step. To generate the contact maps, we aligned the Omni-C data against the corresponding reference with bwa mem [Version 0.7.17-r1188, options -5SP] (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools [Version 0.3.0] (Goloborodko, et al. 2018). We generated a multi-resolution Omni-C matrix with cooler [Version 0.8.10] (Abdennur and Mirny 2020) and balanced it with hicExplorer [Version 3.6] (Ramírez, et al. 2018). We used HiGlass [Version 2.1.11] (Kerpedjiev, et al. 2018) and the PretextSuite (https://github.com/wtsi-hpag/PretextView; https://github.com/wtsi-hpag/PretextMap; https://github.com/wtsi-hpag/PretextSnapshot) to visualize the contact maps.

Using the PacBio HiFi reads and YAGCloser [commit 20e2769] (https://github.com/merlyescalona/yagcloser), we closed some of the remaining gaps generated during scaffolding. We then checked for contamination using the BlobToolKit Framework [Version 2.3.3] (Challis, et al. 2020). Finally, we trimmed remnants of sequence adaptors and mitochondrial contamination.

We assembled the mitochondrial genome of *P. brevispinus* from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (https://github.com/marcelauliano/MitoHiFi) (Allio, et al. 2020). The mitochondrial sequence of *Pisaster ochraceus* (NC_042741.1) was used as the starting reference sequence. After completion of the nuclear genome, we searched for matches of the resulting mitochondrial assembly sequence in the nuclear genome assembly using BLAST+ [Version 2.10] (Camacho, et al. 2009) and filtered out contigs and scaffolds from the nuclear genome with a percentage of sequence identity >99% and size smaller than the mitochondrial assembly sequence.

Nuclear genome size estimation and quality assessment

We generated k-mer counts (k=21) from the PacBio HiFi reads using meryl [Version 1] (https://github.com/marbl/meryl). The generated k-mer database was then used in GenomeScope2.0 [Version 2.0] (Ranallo-Benavidez, et al. 2020) to estimate genome features including genome size, heterozygosity, and repeat content. To obtain general contiguity metrics, we ran QUAST [Version 5.0.2] (Gurevich, et al. 2013). To evaluate genome quality and completeness we used BUSCO [Version 5.0.0] (Simão, et al. 2015; Seppey, et al. 2019) with the metazoan ortholog database (metazoa_odb10) which contains 954 genes. Assessment of base level accuracy (QV) and kmer completeness was performed using the previously generated meryl database and merqury (Rhie, et al. 2020). We further estimated genome assembly accuracy via BUSCO gene set frameshift analysis using the pipeline described in (Korlach, et al. 2017).

We performed a k-means clustering on the lengths of the top 50 *P. brevispinus* scaffolds in R (R Core Team 2013) to test if a drop off in scaffold size corresponded to the number of chromosomes predicted for sea stars (Saotome and Komatsu 2002). We performed k-means clustering on the *P. brevispinus* scaffold lengths in R. The expectation for this test is that longer scaffolds, which represent putative chromosomes, will cluster in one group while shorter scaffolds that were not placed into chromosomes will cluster in a second group based on a measurable change in size between the last putative chromosome scaffold and the first non-chromosome scaffold. The number of long scaffolds in the first cluster therefore gives an estimate of chromosome number in *P. brevispinus*.

Comparison to Pisaster ochraceus genome assembly

We compared the *P. brevispinus* genome assembly produced here to the chromosome-level genome sequence previously published for its congener *P. ochraceus* (Ruiz- Ramos, et al. 2020). We generated completeness metrics for the *P. ochraceus* assembly (ASM1099431v1, GCA_010994315.1) in BUSCO [Version 5.0.0] using the metazoan ortholog database. To determine how the *P. brevispinus* scaffolds correspond to the 22 chromosomes identified in the *P. ochraceus* genome, we

aligned the *P. brevispinus* genome assembly to the *P. ochraceus* chromosomes using the program NUCMER in the MUMmer package [Version 4.0.0] (Marçais, et al. 2018) and visualized the alignments using the program Dot (github.com/marianattestad/dot).

Results

Nucleic acid extraction, library prep, and sequencing

We estimated the integrity of the HMW DNA using the Femto Pulse system and found 96.6% of the DNA fragments were at least 125Kb. The sequencing runs generated 1.1 million PacBio HiFi reads, which yielded ~37-fold coverage (N50 read length 16,677Bp; minimum read length 61Bp; mean read length 16,615Bp; maximum read length of 51,509Bp) based on the Genomescope2.0 genome size estimation of 497.5Mb. Based on the PacBio HiFi reads, we estimated a 0.00238 % sequencing error rate and 1.2% nucleotide heterozygosity rate. The k-mer spectrum output based on the PacBio HiFi reads shows a bimodal distribution with two major peaks, at ~19 and ~38-fold coverage, where peaks correspond to homozygous and heterozygous states respectively of a diploid species.

Nuclear and mitochondrial genome assemblies

We generated a *de novo* nuclear genome assembly for *P. brevispinus* (eaPisBrev1) using PacBio HiFi and Omni-C reads. Complete assembly statistics are reported in Table 2 and Figure 2B. The Omni-C contact maps suggest that both the primary assembly and alternate assemblies are highly contiguous (Figure 2C, Figure S1). The assembled final mitochondrial genome size was 16,223Bp. The base composition of the final assembly version is A=33.09%, C=22.17%, G= 12.91%, T= 31.83%, and consists of 22 unique transfer RNAs and 13 protein coding genes.

Nuclear genome size estimation and quality assessment

Full genome statistics are available in Table 2. The primary assembly consists of 127 scaffolds spanning 505.3Mb with contig N50 of 4.6Mb, scaffold N50 of 21.4Mb, largest contig of 13.9Mb, and largest scaffold of 31.2Mb. The alternate assembly consists of 524 scaffolds spanning 550.1Mb with contig N50 of 3.7Mb, scaffold N50 of 20.4Mb, largest contig of 21.2Mb, and largest scaffold of 34.3Mb. The primary genome assembly size is close to the values from the Genomescope2.0 k-mer spectra (497.5Mb) estimated from the PacBio HiFi reads. The primary assembly has a BUSCO completeness score of 98.7% using the metazoan gene set, a per base quality (QV) of 62.65, a kmer completeness of 84.97 and a frameshift indel QV of 51.94. The alternate assembly has a BUSCO completeness score of 98.7% using the metazoan gene set, a per base quality (QV) of 62.65, a kmer completeness of 84.97 and a frameshift indel QV of 51.94. The k-means clustering analysis placed the longest 23 *P. brevispinus* scaffolds into one group and the remaining scaffolds into a second group (Figure 2D).

Comparison to Pisaster ochraceus *genome assembly*

The *P. brevispinus* genome assembly is ~104Mb larger than *P. ochraceus* (505.3Mb versus 401.9Mb, respectively) and is contained in fewer scaffolds (127 versus 1844, respectively). The scaffold N50 values are similar (21.4Mb and 21.9Mb for *P brevispinus* and *P. ochraceus*, respectively) as was GC content (39.5% and 39.0%, respectively). The *P. brevispinus* assembly has higher BUSCO scores for complete single copy, complete + partial single copy, fragmented, and missing genes, but the *P. ochraceus* genome is superior in number of duplicated genes and the proportion of the genome sequence contained in the largest scaffolds (84% versus 99%, respectively). Whole genome alignment shows that the longest *P. brevispinus* scaffolds generally correspond (blue dots) to the 22 chromosomes predicted for *P. ochraceus*, with one exception — *P. brevispinus* scaffolds 6 and 21 have non-overlapping alignments to *P. ochraceus* chromosome 1, indicating these scaffolds should be joined. Areas of sequence inversion (green dots) and alignment gaps (no dots) were present across the alignment (Figure 3).

Discussion

To generate resources that will inform conservation and management decisions along the California coast and beyond (Shaffer, et al. 2022), we generated a genome assembly for *Pisaster brevispinus*, an ecologically important sea star species. The assembly process we used here (Table 1) aims to generate haplotype-resolved, phased genome assemblies that theoretically correspond to the maternal and paternal chromosomes (Cheng, et al. 2021). Ideally, the two assemblies are similar in size and contiguity, however, variation between assemblies does occur, as we see here for *P. brevispinus* (Table 2). Both the primary and alternate versions of the *P. brevispinus* genome assembly are available on NCBI (Table 2); in the remainder of the discussion, we focus on the more contiguous primary assembly.

Sea star genomes, according to previous karyotyping experiments surveying a range of asteroid species, are organized into 22 chromosomes (Saotome and Komatsu 2002). Recent *de novo* reference genome assembly of *Pisaster ochraceus*, one of two possible sister taxa to *P. brevispinus* (Mah and Foltz 2011), likewise yielded 22 major scaffolds (Ruiz-Ramos, et al. 2020). Our reference genome for *P. brevispinus* is therefore notable in providing a similar yet different estimate, with 23 major scaffolds. Comparison between *P. brevispinus* and *P. ochraceus* shows the source of this difference is that *P. brevispinus* scaffolds 6 and 21 align to Chromosome 1 of *P. ochraceus* (Figure 3). We conclude, therefore, that the genome sequences for *P. ochraceus* (Ruiz-Ramos, et al. 2020) and *P. brevispinus* support the findings of Saotome and Komatsu (2002) — that sea stars have 22 chromosomes (autosomes) — although more data are needed to confirm whether this result is broadly consistent across Asteroidea, or if there is variation in chromosome number in sea stars. Whether asteroids possess a pair of heterotypic, potential sex, chromosomes (Saotome and Komatsu 2002) also remains an open question.

From a DNA sequencing perspective, ideally, the entirety of a genome assembly should be contained within the number of scaffolds equal to the number of actual chromosomes. Moreover, the assembly

should represent the complete genome without gaps or artificial duplication. Comparison of assemblies for the two congeneric sea stars *P. ochraceus* (Ruiz-Ramos, et al. 2020) and *P. brevispinus* (Table 3) provides insight into genome quality beyond that provided by single-genome descriptive statistics, into challenges that remain, and into solutions being offered by recent technological advances. For example, although a higher percentage of genome sequence is contained within the 22 putative sea star chromosomes of *P. ochraceus* (Ruiz-Ramos, et al. 2020) than for *P. brevispinus*, the *P. ochraceus* chromosome sequences include between ~10-20% Ns (Ruiz-Ramos, et al. 2020). The two assemblies also differ in a variety of other aspects including contiguity and completeness (Table 3). Given the congeneric relationship of the taxa, and that the *Pisaster* genomes were generated under similar strategies (i.e., genomic contigs scaffolded with proximity data), differences in the sequencing technologies and assembly algorithms likely explain much of the variation in the assembly statistics. The *P. ochraceus* genome was assembled from Illumina short reads (2 x 150Bp) and scaffolded with Hi-C proximity reads (Ruiz-Ramos, et al. 2020) while the *P. brevispinus* genome was sequenced with PacBio HiFi long reads (mean ~16Kb and max ~51Kb in this study) and scaffolded with Omni-C proximity reads. These differences manifest as lower numbers of fragmented and missing genes, and higher contiguity in *P. brevispinus* compared to *P. ochraceus* because HiFi reads better facilitate assembly through repetitive and low complexity regions relative to short reads. Contiguity in the *P. brevispinus* assembly is also likely increased by the move from scaffolding with Hi-C (which is restriction enzyme specific) to Omni-C (which is restriction enzyme agnostic), which improves resolution of topological interactions in looping and low restriction enzyme regions of the genome. The *P. brevispinus* assembly has higher gene duplication levels than *P. ochraceus*. Older PacBio chemistries had elevated error rates that could lead to artificially increased duplication (Guan, et al. 2020), but *P. brevispinus* was generated with HiFi reads, which have accuracy rates similar to those of Illumina short reads (>99.9%, Dovetail Genomics 2019), which is expected to reduce assembly artifacts. Ultimately, improvements in comparative genomics will require advances to both sequence generation methods (e.g., increasingly long reads) and assembly algorithms (e.g., assembly through repetitive regions and reduced assembly artifacts). Increased taxonomic coverage is also vital for placing the variation we see (e.g., genome size, duplication levels) into a phylogenetic perspective and testing whether these differences represent evolutionary changes between species or technical variation in methods.

The *P. brevispinus* genome generated here is a powerful tool for investigating a range of basic and applied questions central to the California Current Ecosystem. For example, comparative genomics of coastal invertebrates has the potential to further our understanding of local adaptation, connectivity, differentiation, and how these will influence responses to global change. Forthcoming research focused on multiple sea star species will help determine whether areas of structural variation (e.g., sequence inversions, indels, etc.) observed between *P. brevispinus* and *P. ochraceus* (Figure 3) represent assembly artifacts or evolved differences between species, the latter of which have been shown to lead to reproductive isolation and speciation in other marine invertebrate groups (Satou, et al. 2021). Comparison of gene family duplication and loss can explain the evolution of complex traits (Davidson, et al. 2020; Kenny, et al. 2020) and offers a useful strategy for testing genomic drivers of morphological and life history variation across sea stars.

Given the recent rise of mass mortalities, including in *P. brevispinus* and many other sea stars, increasing the number of genome-enabled species will improve the comparative power to test the genetic contribution of a species' susceptibility or tolerance to environmental change and/or disease. For example, previous studies have identified genetic loci responding to selective pressure from sea star wasting in *Pisaster ochraceus* (Schiebelhut et al. 2018, Ruiz-Ramos et al. 2020) and expression differences in loci associated with immune and nervous system function in *Pycnopodia helianthoides* (Fuess, et al. 2015). The availability of reference-quality assemblies will allow us to map such loci to the genome, assign them functional annotations, and compare their sequence and structure, thus permitting important multispecies comparisons and possibly a new perspective on the health of the California Current Ecosystem (CCE).

Multispecies comparisons also will enrich conservation efforts. Genomic data make it possible to better identify biodiversity hotspots and evolutionarily significant units (ESU's) that might require special management (Supple and Shapiro 2018) and can inform captive breeding programs (Hodin, et al. 2021) and efforts to reduce inbreeding depression in depleted populations through assisted gene flow or reintroduction (Frankham 2015; Whiteley, et al. 2015). In line with the goals of the California Conservation Genomics Project (CCGP), we will use the genome as a reference to understand patterns of population genomic structure and demographic change in *P. brevispinus* along the California coast. These data, combined with those for a range of other marine invertebrate taxa also being generated by the CCGP, will provide a comprehensive "community genomics" view of the coast and inform conservation strategies for marine habitats in the CCE.

1. we could amend the end of L355 to read "… multispecies comparisons and possibly a new perspective on ecosystem health of biologically and economically valuable regions, including the California Current Ecosystem" OR "… multispecies comparisons and possibly a new perspective on health of the California Current ecosystem." I prefer the former for setting up a general proposition, and prefer the latter for it's brevity. You could use one, the other, or neither and another.

Funding

This study is a contribution of the Marine Networks Consortium (PIs Michael N Dawson, Rachael A. Bay) as part of the California Conservation Genomics Project (PI: H. Bradley Shaffer), with funding provided to the University of California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224].

Acknowledgements

Thank you to Shannon Myers who spent many dives searching for *P. brevispinus* before finding the specimen used in this study. The California Department of Fish and Wildlife provided the permit under which the specimen was collected (S-200890001-21). PacBio Sequel II library prep and sequencing were carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the Novaseq S4 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. Partial support was provided by Illumina for Omni-C sequencing. Part of this research was conducted using Pinnacles (NSF MRI, # 2019144) at the Cyberinfrastructure and Research Technologies (CIRT) at University of California, Merced. We thank CIRT staff member Robert Romero for computational support.

Data Availability

Data generated for this study are available under NCBI BioProject XXXXXX. Raw sequencing data for sample M0D059179O (NCBI BioSample SAMN26263536) are deposited in the NCBI Short Read Archive (SRA) under XXXXXX for PacBio HiFi sequencing data and XXXXXX for Omni-C Illumina Short read sequencing data. GenBank accessions for both primary and alternate assemblies are XXXXXX and XXXXXX, and for genome sequences XXXXXX and XXXXXX, respectively. The GenBank organelle genome assembly for the mitochondrial genome is XXXXXX. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly

Table 1: Assembly Pipeline and Software Usage. Software citations are listed in the text

Assembly	**Software**	**Version**
Filtering PacBio HiFi adapters	HiFiAdapterFilt https://github.com/sheinasim/HiFiAdapterFilt	Commit 64d1c7b
K-mer counting	Meryl	1
Estimation of genome size and heterozygosity	GenomeScope	2
De novo assembly (contiging)	HiFiasm	0.16.1-r375
Long read, genome-genome alignment	minimap2	2.16
Remove low-coverage, duplicated contigs	purge_dups	1.2.5
Scaffolding		
Omni-C mapping for SALSA	Arima Genomics mapping pipeline https://github.com/ArimaGenomics/mapping_pipeline	Commit 2e74ea4
Omni-C Scaffolding	SALSA	2
Gap closing	YAGCloser https://github.com/merlyescalona/yagcloser	Commit 0e34c3b
Omni-C Contact map generation		
Short-read alignment	bwa	0.7.17-r1188
SAM/BAM processing	samtools	1.11
SAM/BAM filtering	pairtools	0.3.0
Pairs indexing	pairix	0.3.7
Matrix generation	Cooler	0.8.10
Matrix balancing	hicExplorer	3.6

	HiGlass	2.1.11
	PretextMap	0.1.4
	PretextView	0.1.5
Contact map visualization	PretextSnapshot	0.0.3

Table 1 (continued): Assembly Pipeline and Software Usage. Software citations are listed in the text

Organelle assembly		
Mitogenome assembly	MitoHiFi	2 Commit c06ed3e
Genome quality assessment		
Basic assembly metrics	QUAST	5.0.2
Assembly completeness	BUSCO with metazoan gene set	5.0.0
	Merqury	1
Contamination screening		
Local alignment tool	BLAST+	2.10
General contamination screening	BlobToolKit	2.3.3

Table 2: Sequencing and assembly statistics, and accession numbers

Bio Projects & Vouchers	CCGP NCBI BioProject			PRJNA720569
	Genera NCBI BioProject			PRJNA765663
	Species NCBI BioProject			PRJNA808360
	NCBI BioSample			SAMN26263536
	Specimen identification			M0D059179O
	NCBI Genome accessions		**Primary**	**Alternate**
	Assembly accession		pending	pending
	Genome sequences		pending	pending
Genome Sequence	PacBio HiFi reads	Run Accession		
	Omni-C Illumina reads	Run Accession		
Genome Assembly Quality Metrics	Assembly identifier (Quality code *)			eaPisBrev1 (6.6.Q)
	HiFi Read coverage §			37.39X
			Primary	**Alternate**
	Number of contigs		271	694
	Contig N50 (bp)		4,627,265	3,702,536
	Longest Contigs		13,864,835	21,268,271
	Number of scaffolds		127	524
	Scaffold N50 (bp)		21,371,702	20,488,448
	Largest scaffold (bp)		31,152,055	34,371,397
	Size of final assembly (bp)		505,343,882	550,177,770

Gaps per Gbp					269	307
Indel QV (Frame shift)					51.94	51.94
Base pair QV					63.67	62.65
						Full assembly = 63.11
k-mer completeness					83.38	84.97
						Full assembly = 99.14
BUSCO completeness		**C**	**S**	**D**	**F**	**M**
(metazoan) n=954	P‡	98.70%	98.20%	0.50%	0.90%	0.40%
	A‡	98.70%	98.20%	0.50%	0.90%	0.40%
Organelles			1 mitochondrial sequence		Accession number pending	

** Assembly quality code x.y.Q derived notation, from (Rhie et al. 2021). x = log10[contig NG50]; y = log10[scaffold NG50]; Q = Phred base accuracy QV (Quality value). BUSCO Scores. (C)omplete and (S)ingle; (C)omplete and (D)uplicated;(F)ragmented and (M)issing BUSCO genes. n, number of BUSCO genes in the set/database. Bp: base pairs*

§ Read coverage has been calculated based on a genome size of 497Mb.

‡ (P)rimary and (A)lternate assembly values

Table 3 BUSCO metrics for *Pisaster brevispinus* and *Pisaster ochraceus*

	Pisaster brevispinus	*Pisaster ochraceus*
Sequencing technology	PacBio HiFi, Omni-C	Illumina, Hi-C
Assembly length (Mb)	505.3	401.9
Scaffolds	127	1844
Length in predicted chromosomes*	84%	99%
GC-content	39.5%	39.0%
N50 sequence length (Mb)	21.4	21.9
Complete single copy genes	942 (98.7%)	818 (85.7%)
Complete + partial single copy genes	951 (99.7%)	914 (95.8%)
Duplicated genes	5.1 (0.53%)	1.1 (0.12%)
Fragmented genes	8.6 (0.9%)	96.4 (10.1%)
Missing genes	3.8 (0.4%)	40.1 (4.2%)

*longest 22 scaffolds for *P. ochraceus* and longest 23 scaffolds for *P. brevispinus*, see Results, Discussion

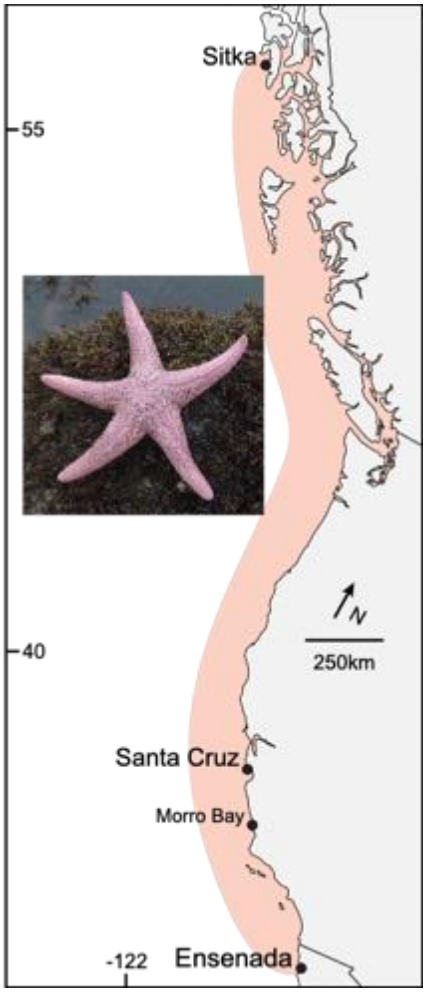

Figure 1. Map showing the current known distribution (ca. 2022) of the giant pink sea star, *Pisaster brevispinus*, from Sitka, Alaska, USA to Ensenada, Baja California, Mexico. The individual sequenced for this study was collected at Terrace Point, Santa Cruz, California on 13 October 2020. The inset shows a photograph of a *P. brevispinus* sea star at Morro Bay, California. Photo credit: Jerry Kirkhart (https://www.flickr.com/photos/jkirkhart35/423749561/), some rights reserved (https://creativecommons.org/licenses/by/4.0)

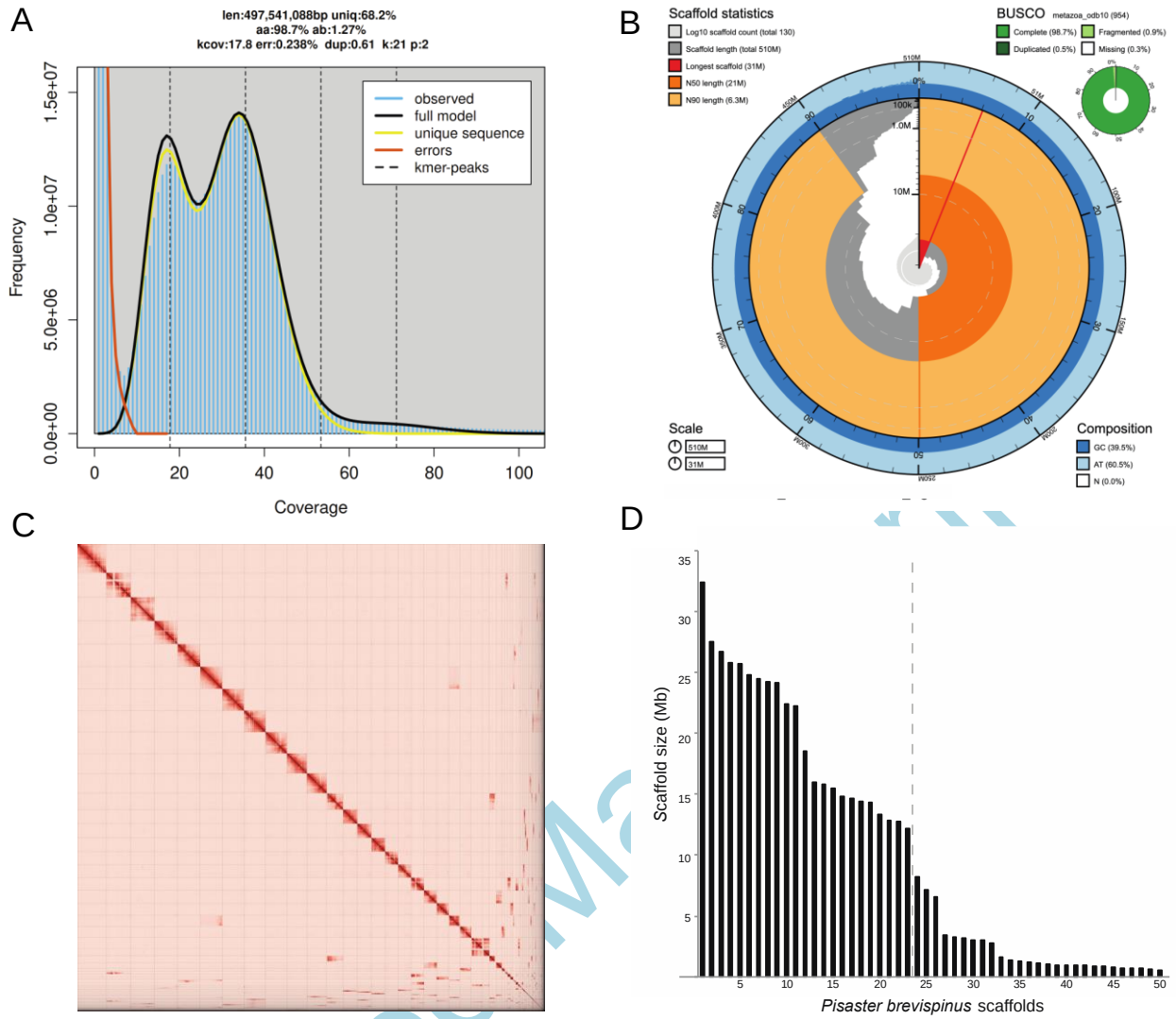

Figure 2. Visual overview of genome assembly metrics. (A) K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage and lower frequency correspond to differences between haplotypes, whereas the higher coverage and frequency k-mers correspond to the similarities between haplotypes. (B) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *Pisaster brevispinus* primary assembly (eaPisBrev1). The plot circle represents the full size of the assembly. From the inside-out, the central plot covers length-related metrics. The red line represents the size of the longest scaffold; all other scaffolds are arranged in size-order moving clockwise around the plot and drawn in gray starting from the outside of the central plot. Dark and light orange arcs show the scaffold N50 and scaffold N90 values. The central light gray spiral shows the cumulative scaffold count with a white line at each order of magnitude. White regions in this area reflect the proportion of Ns in the assembly The dark vs. light blue area around it shows mean, maximum and minimum GC vs. AT content at 0.1% intervals. (C) Omni-C Contact maps for the primary genome assembly generated with PretextSnapshot. Omni-C contact maps translate proximity of genomic regions in 3-D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two such regions. Scaffolds are separated by black lines and higher density corresponds to higher levels of fragmentation. (D) Histogram of the 50 largest *P. brevispinus* scaffolds. Gray dashed line represents the break point for two clusters delimited by k-means clustering of scaffold lengths.

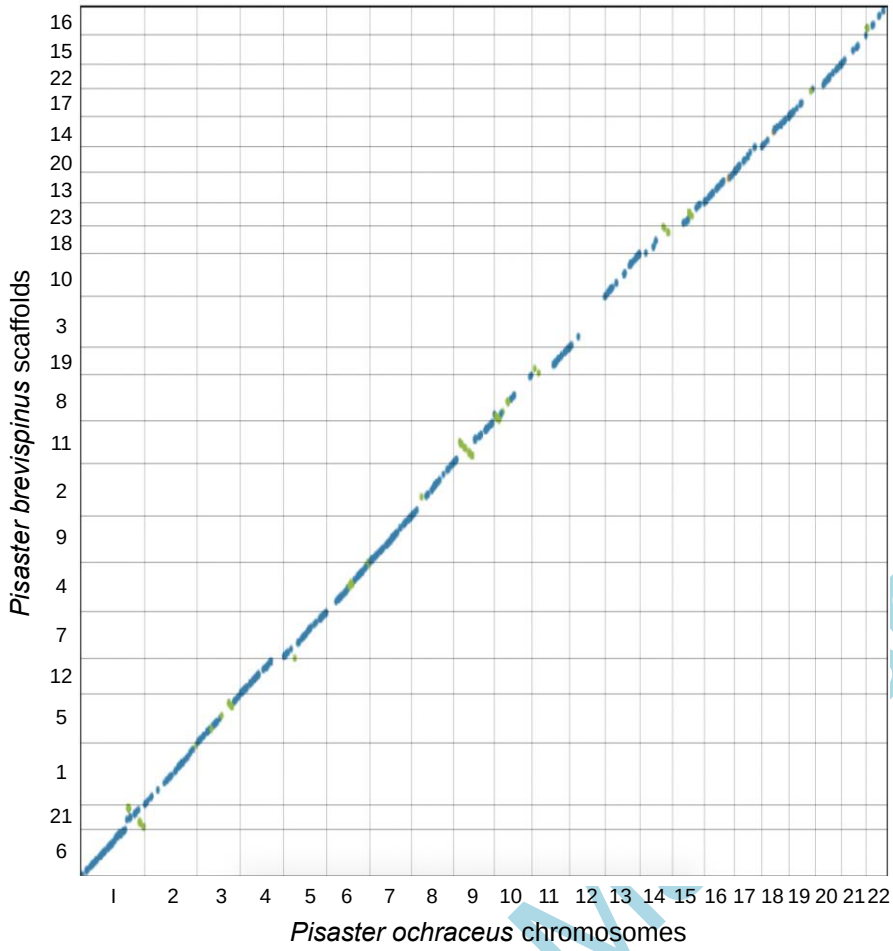

Figure 3. Whole genome alignment between the predicted *Pisaster ochraceus* 22 chromosomes (x-axis) and top 23 longest *Pisaster brevispinus* primary assembly scaffolds (y-axis). Blue dots represent areas of sequence alignment in the same direction and green dots represent areas of inverted sequence alignment in *P. brevispinus* (the query) relative to the *P. ochraceus* sequence (the reference). Light gray lines indicate chromosome and scaffold boundaries. The total axes are scaled by sequence length contained in top 23 *P. brevispinus* scaffolds (437.4Mb) and *P. ochraceus* chromosomes (398.1Mb). Each scaffold-to-chromosome alignment block is scaled by the length of the *P. ochraceus* chromosome (x-axis) and the *P. brevispinus* scaffold (y-axis).

References

Abdennur N, Mirny LA. 2020. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics 36:311-316.

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsuc F. 2020. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources 20:892-905.

Beas-Luna R, Micheli F, Woodson CB, Carr M, Malone D, Torre J, Boch C, Caselle JE, Edwards M, Freiwald J. 2020. Geographic variation in responses of kelp forest communities of the California Current to recent climatic changes. Global Change Biology 26:6457-6473.

Blanchette CA, Melissa Miner C, Raimondi PT, Lohse D, Heady KE, Broitman BR. 2008. Biogeographical patterns of rocky intertidal communities along the Pacific coast of North America. Journal of Biogeography 35:1593-1607.

Burt JM, Tinker MT, Okamoto DK, Demes KW, Holmes K, Salomon AK. 2018. Sudden collapse of a mesopredator reveals its complementary role in mediating rocky reef regime shifts. Proceedings of the Royal Society B: Biological Sciences 285:20180553.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:1-9.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit–interactive quality assessment of genome assemblies. G3: Genes, Genomes, Genetics 10:1361-1374.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nature Methods 18:170-175.

Connell JH. 1972. Community interactions on marine rocky intertidal shores. Annual review of Ecology and Systematics:169-192.

Costello MJ, Bouchet P, Boxshall G, Fauchald K, Gordon D, Hoeksema BW, Poore GC, van Soest RW, Stöhr S, Walter TC. 2013. Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases. PloS One 8:e51629.

Davidson PL, Guo H, Wang L, Berrio A, Zhang H, Chang Y, Soborowski AL, McClay DR, Fan G, Wray GA. 2020. Chromosomal-level genome assembly of the sea urchin *Lytechinus variegatus* substantially improves functional genomic analyses. Genome Biology and Evolution 12:1080-1086.

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Höglund J. 2022. The era of reference genomes in conservation genomics. Trends in Ecology & Evolution.

Frankham R. 2015. Genetic rescue of small inbred populations: Meta-analysis reveals large and consistent benefits of gene flow. Molecular Ecology 24:2610-2618.

Fuess LE, Eisenlord ME, Closek CJ, Tracy AM, Mauntz R, Gignoux-Wolfsohn S, Moritsch MM, Yoshioka R, Burge CA, Harvell CD. 2015. Up in arms: immune and nervous system response to sea star wasting disease. PLoS One 10:e0133053.

Genomics D. 2019. https://dovetailgenomics.com/wp-content/uploads/2019/08/Omni-C_TechNote.pdf.

Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. 2017. Scaffolding of long read assemblies using long range contact information. BMC Genomics 18:1-11.

Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillippy AM, Koren S. 2019. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Computational Biology 15:e1007273.

Goloborodko A, Abdennur N, Venev S, Brandao H, Fudenberg G. 2018. mirnylab/pairtools: v0.2.0. 10.5281/zenodo.1490831.
Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics 36:2896-2898.
Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072-1075.
Harvell CD, Lamb JB. 2020. Disease outbreaks can threaten marine biodiversity. Marine Disease Ecology; Oxford University Press: Oxford, UK:141-158.
Hodin J, Pearson-Lund A, Anteau F, Kitaeff P, Cefalu S. 2021. Progress toward complete life-cycle culturing of the endangered sunflower star, *Pycnopodia helianthoides*. The Biological Bulletin 241:243-258.
Jurgens LJ, Rogers-Bennett L, Raimondi PT, Schiebelhut LM, Dawson MN, Grosberg RK, Gaylord B. 2015. Patterns of mass mortality among rocky shore invertebrates across 100 km of northeastern Pacific coastline. PloS One 10:e0126280.
Kenny NJ, Francis WR, Rivera-Vicéns RE, Juravel K, de Mendoza A, Díez-Vives C, Lister R, Bezares-Calderón LA, Grombacher L, Roller M. 2020. Tracing animal genomic evolution with the chromosomal-level assembly of the freshwater sponge *Ephydatia muelleri*. Nature Communications 11:1-11.
Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N. 2018. HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biology 19:1-12.
Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Cantin L, Jarvis ED. 2017. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. Gigascience 6:gix085.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
Lin M, Escalona M, Sahasrabudhe R, Nguyen O, Beraut E, Buchalski MR, Wayne RK. 2022. A Reference Genome Assembly of the Bobcat, Lynx rufus. Journal of Heredity.
Mah C, Foltz D. 2011. Molecular phylogeny of the Forcipulatacea (Asteroidea: Echinodermata): systematics and biogeography. Zoological Journal of the Linnean Society 162:646-660.
Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. PLoS Computational Biology 14:e1005944.
McPherson ML, Finger DJ, Houskeeper HF, Bell TW, Carr MH, Rogers-Bennett L, Kudela RM. 2021. Large-scale shift in the structure of a kelp forest ecosystem co-occurs with an epizootic and marine heatwave. Communications Biology 4:1-9.
Menge BA, Caselle JE, Barth JA, Blanchette CA, Carr MH, Chan F, Gravem S, Gouhier TC, Lubchenco J, McManus MA. 2019. Community responses to climate-related variability and disease. Oceanography 32:72-81.
Montecino-Latorre D, Eisenlord ME, Turner M, Yoshioka R, Harvell CD, Pattengill-Semmens CV, Nichols JD, Gaydos JK. 2016. Devastating transboundary impacts of sea star wasting disease on subtidal asteroids. PloS One 11:e0163190.
Morris RH, Abbott DP, Haderlie EC. 1980. Intertidal invertebrates of California: Stanford University Press Stanford.
Paine RT. 1966. Food web complexity and species diversity. The American Naturalist 100:65-75.

Ramírez F, Bhardwaj V, Arrigoni L, Lam KC, Grüning BA, Villaveces J, Habermann B, Akhtar A, Manke T. 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. Nature Communications 9:1-15.
Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications 11:1-10.
Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology 21:1-27.
Ruiz-Ramos DV, Schiebelhut LM, Hoff KJ, Wares JP, Dawson MN. 2020. An initial comparative genomic autopsy of wasting disease in sea stars. Molecular Ecology 29:1087-1102.
Sagarin RD, Barry JP, Gilman SE, Baxter CH. 1999. Climate-related change in an intertidal community over short and long time scales. Ecological Monographs 69:465-490.
Sanford E, Sones JL, García-Reyes M, Goddard JH, Largier JL. 2019. Widespread shifts in the coastal biota of northern California during the 2014–2016 marine heatwaves. Scientific Reports 9:1-14.
Saotome K, Komatsu M. 2002. Chromosomes of Japanese starfishes. Zoological Science 19:1095-1103.
Satou Y, Sato A, Yasuo H, Mihirogi Y, Bishop J, Fujie M, Kawamitsu M, Hisata K, Satoh N. 2021. Chromosomal inversion polymorphisms in two sympatric ascidian lineages. Genome Biology and Evolution 13:evab068.
Schultz JA, Cloutier RN, Côté IM. 2016. Evidence for a trophic cascade on rocky reefs following sea star mass mortality in British Columbia. PeerJ 4:e1980.
Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. In. Gene prediction: Springer. p. 227-245.
Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Gold M, Harrigan RJ, Hodges S, Luckau TK. 2022. Landscape genomics to enable conservation actions: the California Conservation Genomics Project. Journal of Heredity.
Sim SB, Corpuz RL, Simmonds TJ, Geib SM. 2022. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. BMC Genomics 23:1-7.
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210-3212.
Strathmann R. 1987. Larval feeding. Reproduction of marine invertebrates 9:465-550.
Supple MA, Shapiro B. 2018. Conservation of biodiversity in the genomics era. Genome Biology 19:1-12.
Team RC. 2013. R: A language and environment for statistical computing.
Weber ED, Auth TD, Baumann-Pickering S, Baumgartner TR, Bjorkstedt EP, Bograd SJ, Burke BJ, Cadena-Ramírez JL, Daly EA, de la Cruz M. 2021. State of the California Current 2019–2020: Back to the Future With Marine Heatwaves? Frontiers in Marine Science:1081.
Whiteley AR, Fitzpatrick SW, Funk WC, Tallmon DA. 2015. Genetic rescue to the rescue. Trends in Ecology & Evolution 30:42-49.