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

Melissa B. DeBiasse¹, Lauren M. Schiebelhut¹, Merly Escalona², Eric Beraut³, Colin Fairbairn³, Mohan P.A. Marimuthu⁴, Oanh Nguyen⁴, Ruta Sahasrabudhe⁴, Michael N Dawson¹

¹Department of Life and Environmental Sciences, University of California, Merced, CA, United States, ²Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, United States, ³Ecology & Evolutionary Biology Department, University of California Santa Cruz, Santa Cruz, CA, United States, ⁴DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California Davis, Davis, CA, United States

Address correspondence to M.B. DeBiasse at the address above, or e-mail: melissa.debiasse@gmail.com.

Corresponding Editor: Jose Lopez

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 550 Mb in length. The assembly contains 127 scaffolds with a contig N50 of 4.6 Mb and a scaffold N50 of 21.4 Mb; the BUSCO completeness score is 98.70%. The *P. brevispinus* genome assembly is comparable to the genome of the congener 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.

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

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 have 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 2 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, Pismaster ochraceus, in the Northeast Pacific (Paine 1966; Schultz et al. 2016). Of the 20 or more species impacted by the geographically and phylogenetically broad sea star wasting outbreak in 2013, Pismaster 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). 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, United States (Fig. 1, Morris et al. 1980; Costello et al. 2013; Beasley-Luna et al. 2020). They are gonochoristic broadcast spawners (Morris et al. 1980) with an estimated larval duration of 76 to 266 d (Strathmann 1987). P. 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 ~16 cm), 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 sea stars 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, 118 mm radius (arm tip to disc center), was collected from a sandstone platform at 11 to 13 m depth at Terrace Point, Santa Cruz, CA, United States (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, United States.

Nucleic acid extraction, library preparation, and sequencing

We extracted high molecular weight (HMW) genomic DNA (gDNA) from 28 mg 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 18,000 × g (instead of recommended 1,500 × g) 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 (210 ng/µl; 20 µg total) using the Quantus Fluorometer and checked the size distribution using the Agilent Femto Pulse. We concentrated the sheared gDNA using 0.45× of AMPure PB beads followed by quantification using a Quantus Fluorometer. We used 6 µg of sheared, concentrated DNA as input for the removal of single-strand overhangs at 37 °C for 15 min, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 min, end repair and A-tailing at 20 °C for 10 min and 65 °C for 30 min, ligation of overhang adapter v3 at 20 °C.
for 1 h and 65 °C for 10 min to inactivate the ligase, and nuclease treatment of SMRTbell library at 37 °C for 1 h to remove damaged or nonintact SMRTbell templates. We purified and concentrated the SMRTbell library with 0.8× AMPure PB beads for size selection using the BluePippin system. We purified the input of 3.2 µg purified SMRTbell library to load into the Blue Pippin 0.75% Agarose Cassette using cassette definition 0.75% DF Marer S1 3 to 10 kb Improved Recovery for the run protocol. We collected fragments greater than 7 kb from the cassette elution well and purified and concentrated the size-selected SMRTbell library with 0.8× 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 100 and 40 µm 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 γ-adaptor using the NEB Ultra II DNA Library Prep kit and captured biotin-containing fragments using streptavidin beads. We split the postcapture product into 2 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-h movies on a PacBio Sequel II sequencer. The Omni-C library was sequenced on an Illumina NovaSeq platform to generate approximately 100 million 2 × 150 bp 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. 2021) 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://jh3.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, 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 -SSP] (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools [Version 0.3.0] (Golobrodko et al. 2018). We generated a multiresolution Omni-C matrix with cooler [Version 0.8.10] (Abdenur 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 adapters 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 P. 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 GenomeScope 2.0 [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. 2013; Seppey et al. 2019) with the metazoan ortholog database (metazoa_odb10) which contains 954 genes. Assessment of base level accuracy (QV) and k-mer completeness was performed using the previously generated meryl database and merqury (Rhide 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 2022) to test if a drop off in scaffold size corresponded to the number of chromosomes predicted for sea stars (Saotome and Komatsu 2002). The expectation for this test is that longer scaffolds, which represent putative chromosomes, will cluster in a 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 nonchromosome
scaffold. The number of long scaffolds in the first cluster therefore gives an estimate of chromosome number in *P. brevispinus*.

Comparison to *P. 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 125 kb. The sequencing runs generated 1.1 million PacBio HiFi reads, which yielded ~37-fold coverage (N50 read length 16,677 bp; minimum read length 61 bp; mean read length 16,615 bp; maximum read length 51,509 bp) based on the GenomeScope 2.0 genome size estimation of 497.5 Mb. 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 2 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 Fig. 2B. The Omni-C contact maps suggest that both the primary assembly and alternate assemblies are highly contiguous (Fig. 2C, Supplementary Fig. S1). The assembled final mitochondrial genome size was 16,223 bp. 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 505.3 Mb with contig N50 of 4.6 Mb, scaffold N50 of 21.4 Mb, largest contig of 13.9 Mb, and largest scaffold of 31.2 Mb. The alternate assembly consists of 524 scaffolds spanning 550.1 Mb with contig N50 of 3.7 Mb, scaffold N50 of 20.4 Mb, largest contig of 21.2 Mb, and largest scaffold of 34.3 Mb. The

---

**Table 2. Sequencing and assembly statistics, and accession numbers.**

| BioProjects and vouchers | CCGP NCBI BioProject | PRJNA720569 |
|--------------------------|----------------------|-------------|
| Genera NCBI BioProject   | PRJNA765663          |             |
| Species NCBI BioProject  | PRJNA808360          |             |
| NCBI BioSample           | SAMN26263536         |             |
| Specimen identification  | M0D059179O           |             |
| NCBI Genome accessions   |                      |             |
| Assembly accession       | PRJNA810506          | PRJNA810505 |
| Genome sequences         | GCA_023634235.1      | GCA_023634265.1 |

| Genome sequence          | PacBio HiFi reads     | SRR19909216 |
|--------------------------|-----------------------|-------------|
|                          | SRX15951960           |             |
| Omni-C Illumina reads    | SRR19909214-SRR19909215 |            |
|                          | SRX15951961-SRX15951962 |         |

| Genome Assembly Quality Metrics | Assembly identifier (quality code) | eaPisBrev1 (6.6.Q) |
|---------------------------------|-----------------------------------|-------------------|
| HiFi read coverage              |                                   | 37.39×            |

| Assembly identifier (quality code) | eaPisBrev1 (6.6.Q) |
|-----------------------------------|-------------------|
| HiFi read coverage                | 37.39×            |

| 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 (frameshift) | 51.94 | 51.94 |
| Base pair QV       | 63.67 | 62.65 |

| K-mer completeness | 83.38 | 84.97 |
| BUSCO completeness (metazoan), n = 954 |
| C                  | 98.70% | 98.20% | 0.50% | 0.90% | 0.40% |
| S                  | 98.70% | 98.20% | 0.50% | 0.90% | 0.40% |
| D                  | 98.70% | 98.20% | 0.50% | 0.90% | 0.40% |
| F                  | 98.70% | 98.20% | 0.50% | 0.90% | 0.40% |
| M                  | 98.70% | 98.20% | 0.50% | 0.90% | 0.40% |

| Organelles | 1 mitochondrial sequence | JAMKCF010000513.1 |
|------------|--------------------------|-------------------|

---

*Assembly quality code x:y:Q derived notation, from Rhie et al. (2022). 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.

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

*(P)rimary and (A)lternate assembly values.
primary genome assembly size is close to the values from the GenomeScope 2.0 k-mer spectra (497.5 Mb) 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 k-mer 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 a group and the remaining scaffolds into a second group (Fig. 2D).

**Comparison to *P. ochraceus* genome assembly**

The *P. brevispinus* genome assembly is ~104 Mb larger than *P. ochraceus* (505.3 vs. 401.9 Mb, respectively) and is contained in fewer scaffolds (127 vs. 1,844, respectively). The scaffold N50 values are similar (21.4 and 21.9 Mb for *P. brevispinus* and *P. ochraceus*, respectively) as was GC content (39.5% and 39.0%, respectively). The *P. brevispinus* scaffolds into a group and the remaining scaffolds into a second group (Fig. 2D).
assembly has higher BUSCO scores for complete single copy, complete + partial single copy, fragmented, and missing genes, but the \textit{P. ochraceus} genome is superior in number of duplicated genes and the proportion of the genome sequence contained in the largest scaffolds (84\% vs. 99\%, respectively). Whole genome alignment showed that the longest \textit{P. brevispinus} scaffolds generally correspond (blue dots) to the 22 chromosomes predicted for \textit{P. ochraceus}, with one exception—\textit{P. brevispinus} scaffolds 6 and 21 have nonoverlapping alignments to \textit{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 (Fig. 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 \textit{P. 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 2 assemblies are similar in size and contiguity, however, variation between assemblies does occur, as we see here for \textit{P. brevispinus} (Table 2). Both the primary and alternate versions of the \textit{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 \textit{P. ochraceus}, 1 of 2 possible sister taxa to \textit{P. brevispinus} (Mah and Foltz 2011), likewise yielded 22 major scaffolds (Ruiz-Ramos et al. 2020). Our reference genome for \textit{P. brevispinus} is therefore notable in providing a similar yet different estimate, with 23 major scaffolds. Comparison between \textit{P. brevispinus} and \textit{P. ochraceus} shows the source of this difference is that \textit{P.
Table 3. BUSCO metrics for Pisaster brevispinus and Pisaster ochraceus.

| Metric                  | Pisaster brevispinus | Pisaster ochraceus |
|------------------------|----------------------|--------------------|
| Sequencing technology  | PacBio HiFi, Omni-C  | Illumina, Hi-C     |
| Assembly length (Mb)   | 505.3                | 401.9              |
| Scaffolds              | 127                  | 1,844              |
| 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.

brevispinus scaffolds 6 and 21 align to chromosome 1 of P. ochraceus (Fig. 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 the 2 assemblies for the 2 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% and 20% Ns (Ruiz-Ramos et al. 2020). The 2 assemblies also differ in a variety of other aspects including contiguousity 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 × 150 bp) 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 ~16 kb and max ~51 kb in this study) and scaffolded with Omni-C proximity reads. These differences manifest as lower numbers of fragmented and missing genes, and higher contiguousity in P. brevispinus compared with 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%, https://dovetailgenomics.com/wp-content/uploads/2019/08/Omni-C_TechNote.pdf), 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 CCE. 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 (Fig. 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 P. ochraceus (Schiebelhut et al. 2018; Ruiz-Ramos et al. 2020) and expression differences in loci associated with immune and nervous system function in Pyrocopida belantiboides (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 CCE.

Multispecies comparisons also will enrich conservation efforts. Genomic data make it possible to better identify biodiversity hotspots and evolutionarily significant units 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 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.

**Supplementary material**

Supplementary material is available at *Journal of Heredity* online.

Fig. S1. Omni-C contact map for the alternate genome assembly generated with PretextSnapshot. Omni-C contact maps translate proximity of genomic regions in 3D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between 2 such regions. Scaffolds are separated by black lines and higher density corresponds to higher levels of fragmentation.

**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].

**Acknowledgments**

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 Core 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

PRJNA808360. Raw sequencing data for sample M0D0591790 (NCBI BioSample SAMN26263536) are deposited in the NCBI Short Read Archive (SRA) under SRX15951960 for PacBio HiFi sequencing data and SRX15951961-SRX15951962 for Omni-C Illumina Short read sequencing data. GenBank accessions for both primary and alternate assemblies are PRJNA810506 and PRJNA810505, and for genome sequences GCA_023634265.1 and GCA_023634265.1, respectively. The GenBank organelle genome assembly for the mitochondrial genome is XXXXXXX. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly.

**References**

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

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, Delsec F. MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. *Mol Ecol Resour*. 2020;20:892–905.

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

Blanchette CA, Melissa Miner C, Raimondi PT, Lohse D, Heady KE, Brostman BR. Biogeographical patterns of rocky intertidal communities along the Pacific coast of North America. *J Biogeogr*. 2008;35:1593–1607.

Burt JM, Tinker MT, Okamoto DK, Demes KW, Holmes K, Salomon AK. Sudden collapse of a mesopredator reveals its complementary role in mediating rocky reef regime shifts. *Proc R Soc B Biol Sci*. 2018;285:20180553.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Beaker K, Madden TL. BLAST+: architecture and applications. *BMC Bioinf*. 2009;10:1–9.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3*. 2020;10:1361–1374.

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

Connell JH. Community interactions on marine rocky intertidal shores. *Annu Rev Ecol Syst*. 1972;3:169–192.

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

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

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, Ciofi C, Crottini A, Godoy JA, Hoflund J. The era of reference genomes in conservation genomics. *Trends Ecol Evol*. 2022;37:197–202.

Frankham R. Genetic rescue of small inbred populations: meta-analysis reveals large and consistent benefits of gene flow. *Mol Ecol*. 2015;24:2610–2618.

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

Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffolding of long read assemblies using phased assembly graphs with hifiasm. *Nat Methods*. 2017;18:1–11.

Ghurye J, Rhee A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Phillips AM, Koren S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput Biol*. 2019;15:e1007273.

Goloborodko A, Abdennur N, Venev S, Brandao H, Fudenberg G. mirnylab/pairtools: v0.2.0. 2018. doi:10.5281/zenodo.1490831
Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*. 2020;36:2896–2898.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29:1072–1075.

Harvell CD, Lamb JB. Disease outbreaks can threaten marine biodiversity. *Oxford (UK): Marine Disease Ecology, Oxford University Press*. 2020. p. 141–158.

Hodin J, Pearson-Lund A, Anteau E, Kitaeff P, Cefalu S. Progress to ward complete life-culture of the endangered sunflower star, *Pseudonodidae solanoides*. *Bioll Bull*. 2021;241:243–258.

Jürgens LJ, Rogers-Bennett I, Raimondi PT, Schiebelhut LM, Dawson MN, Grosberg RK, Gaylord B. Patterns of mass mortality among rocky shore invertebrates across 100 km of northeastern Pacific coastline. *PLoS One*. 2015;10:e0126280.

Kenny NJ, Francis WR, Rivera-Vicins RE, Juravel K, de Mendoza A, Hodin J, Pearson-Lund A, Anteau F, Kitaeff P, Cefalu S. Progress to ward complete life-culture of the endangered sunflower star, *Pseudonodidae solanoides*. *Bioll Bull*. 2021;241:243–258.

Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Shaffer HB, Toffelmier E, Corbett-Detig RB, Escalona M, Sahasrabudhe R, Nguyen O, Beraut E, Buchalski M, Foltz D. Molecular phylogeny of the Forcipulatacea (Asteroidea: Echinodermata): systematics and biogeography. *Zool J Linn Soc*. 2011;162:646–660.

Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, *arXiv*. arXiv:1303.3997, 2013, preprint: not peer reviewed. doi:10.48550/arXiv.1303.3997

Lin M, Escalona M, Sahasrabudhe R, Nguyen O, Beraut E, Buchalski MR, Wayne RK. A Reference genome assembly of the bobcat, *Lynx rufus*. *J Hered*. 2022;11:615–623.

Mah C, Foltz D. Molecular phylogeny of the Forcipulatacea (Asteroidea: Echinodermata): systematics and biogeography. *Zool J Linn Soc*. 2011;162:646–660.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol*. 2018;14:e1005944.

McPherson ML, Finger DJ, Houskeeper HF, Bell TW, Carr MH, Rogers-Bennett I, Kudela RM. Large-scale shift in the structure of a kelp forest ecosystem co-occurs with an epizootic and marine heatwave. *Commun Biol*. 2021;4:1–9.

Menge BA, Caselle JE, Barth JA, Blanchette CA, Carr MH, Chan F, Greaves S, Gouhier TC, Lubchenco J, McNamans MA. Community responses to climate-related variability and disease. *Oceanography*. 2019;32:72–81.

Montecino-Latorre D, Eisenlord ME, Turner M, Yoshioka R, Harvell CD, Pattengill-Semmens CV, Nichols JD, Gaydos JK. Devastating transboundary impacts of sea star wasting disease on subtidal asteroids. *PLoS One*. 2016;11:e0163190.

Morris RH, Abbott DP, Haderle EC. *Intertidal invertebrates of California*. Stanford California USA; Stanford University Press Stanford; 1980.

Paine RT. Food web complexity and species diversity. *Am Nat*. 1966;100:65–75.