Reference genome assembly of the sunburst anemone, *Anthopleura sola*

**Brendan H. Cornwell**1,2, **Eric Beraut**2, **Colin Fairbairn**2, **Oanh Nguyen**3,10, **Mohan P.A. Marimuthu**4, **Merly Escalona**4, **Erin Toffelmier**5,6,10

1Department of Biology, Hopkins Marine Station of Stanford University, Pacific Grove, CA, United States,
2Department of Ecology and Evolutionary Biology, University of California, Santa Cruz, Santa Cruz, CA, United States,
3DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California, Davis, CA, United States,
4Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, United States,
5Department of Ecology & Evolutionary Biology, University of California, Los Angeles, CA, United States,
6La Kretz Center for California Conservation Science, Institute of the Environment and Sustainability, University of California, Los Angeles, CA, United States

Address correspondence to B.H. Cornwell at the address above, or e-mail: bhcornwell@ucdavis.edu.

Corresponding Editor: Kenneth Olsen

**Abstract**

The sunburst anemone *Anthopleura sola* is an abundant species inhabiting the intertidal zone of coastal California. Historically, this species has extended from Baja California, Mexico to as far north as Monterey Bay, CA. However, recently the geographic range of this species has expanded to Bodega Bay, CA, possibly as far north as Salt Point, CA. This species also forms symbiotic partnerships with the dinoflagellate *Breviolum muscatinei*, a member of the family Symbiodiniaceae. These partnerships are analogous to those formed between tropical corals and dinoflagellate symbionts, making *A. sola* an excellent model system to explore how hosts will (co)evolve with novel symbiont populations they encounter as they expand northward. This assembly will serve as the foundation for identifying the population genomic patterns associated with range expansions, and will facilitate future work investigating how hosts and their symbiont partners will evolve to interact with one another as geographic ranges shift due to climate change.

**Key words:** *Anthopleura sola*, California Conservation Genomics Project, CCGP, range expansion, symbiosis

**Introduction**

The sunburst anemone, *Anthopleura sola* is a large, solitary anemone inhabiting the intertidal zone of the Pacific coast from Baja California, Mexico to central California (Fig. 1). Within the past half century, the geographic range of this species has expanded northward (Denny and Gaines 2007), likely ending between Bodega Bay, CA and Salt Point (Mendocino County), CA (BHC, pers. obs.). Northward expansions of species historically relegated to more equatorial latitudes along the California coast have been documented during temporary periods of increased temperatures near geographic range edges (Sanford et al. 2019), but have nonetheless prompted researchers to begin assessing how these populations will evolve to match novel geographic locations they encounter during longer-term range expansions. A particularly important feature of the *A. sola* expansion is that they are likely encountering novel symbiotic populations that historically have only interacted with 2 other symbiotic species that are members of the genus *Anthopleura*, *A. xanthogrammica* and *A. elegantissima* whose geographic ranges extend to Alaska. Previous work has shown that these symbionts are shared between these 3 species—-with the exception of the southernmost populations of *A. sola* and *A. elegantissima*, where symbionts are partitioned by host species (Cornwell and Hernández 2021). As *A. sola* continues to move northward, interactions between newly arriving hosts and naive symbiont populations will become more common, which will allow researchers to identify patterns of molecular and physiological coevolution in both partners as geographic ranges shift with climate change. Because this symbiosis is analogous to the partnership between tropical corals and dinoflagellates, characterizing how novel symbiotic partnerships evolve along the California coast will have global implications.

Well-assembled genomes are an important tool for identifying genomic patterns associated with range expansions and coevolution with symbiont partners. *A. sola* exhibits little population structure across its geographic range with no evidence for historical population bottlenecks, which likely contributes to the highest average level of heterozygosity of the 3 symbiotic species of *Anthopleura* on the Pacific coast of North America (τ = 0.0095; Cornwell and Hernández 2021). A draft assembly of *A. sola* has already been generated...
without long reads, resulting in a contig N50 of 5,224 bp and scaffold N50 of 16,096 with a total estimated genome size of 434 Mb (Cornwell 2020). Here, we present a new assembly for *A. sola* which substantially improves on previously published versions, and creates a new resource for marine scientists studying how marine populations will evolve as their geographic ranges shift with warming conditions.

**Methods**

**Biological materials**

DNA was extracted from a single *A. sola* polyp collected in Pacific Grove, CA by Brendan Cornwell (36.621707, −121.904580). All tissue for sequencing was preserved by snap freezing in liquid nitrogen and shipping on dry ice for the University of California, Davis and the University of California, Santa Cruz, the remaining tissue was preserved in ethanol (Fig. 1).

**Nucleic acid library preparation and sequencing**

High molecular weight (HMW) genomic DNA (gDNA) was extracted from 60 mg of body wall tissue using the Nanobind Tissue Big DNA kit as per the manufacturer’s instructions (Pacific BioSciences—PacBio, Menlo Park, CA) with the following modifications. After the second resuspension step, we pelleted the tissue homogenate by centrifuging at 16,000 × g (4 °C for 5 min) to remove the residual wash buffer and performed the lysis step with 1.5× reaction volume. The DNA purity was estimated using absorbance ratios (260/280 = 1.84 and 260/230 = 2.28) on the NanoDrop ND-1000 spectrophotometer. The final DNA yield (334 ng/µL; 32 µg) was quantified using the Quantus Fluorometer (QuantiFluor ONE dsDNA Dye assay, Promega, Madison, WI). The size distribution of the HMW DNA was estimated using the Femto Pulse system (Agilent, Santa Clara, CA) and found that 70% of the fragments were 120 kb.

We generated long reads for the assembly using the SMRTbell Express Template Prep Kit v2.0 (PacBio, Cat. #100-938-900). Briefly, HMW DNA was sheared to 15 to 20 kb, cleaned and end-repaired, and finally size selected using the BluePippin system (Sage Science, Beverly, MA; Cat. #BLF7510) to generate a library of fragments greater than 9 kb. The 15 to 20 kb average HiFi SMRTbell library was sequenced at UC Davis DNA Technologies Core (Davis, CA) using 3 SMRT cells, Sequel II sequencing chemistry 2.0, and 30-h movies each on a PacBio Sequel II sequencer.

We prepared Omni-C libraries using the Dovertail Omni-C Kit (Dovetail Genomics, Scotts Valley, CA). After grinding tissue with a mortar and pestle under liquid nitrogen, chromatin was fixed in the nucleus and strained through 100 and 40 µm cell strainers. We digested the chromatin using DNaseI to generate Illumina-compatible length distributions, purified the DNA and generated an NGS library using an NEB Ultra II DNA Library Prep kit (New England Biolabs, Ipswich, MA). The library was sequenced at Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq platform (Illumina, San Diego, CA) to generate approximately 100 million 2 × 150 bp read pairs per GB of genome size.

**Nuclear genome assembly**

We assembled the genome of *A. sola* following the CCGP assembly protocol Version 4.0, which produces a high-quality and highly contiguous assembly using PacBio HiFi reads and Omni-C data while minimizing manual curation (outlined on Table 1). Briefly, we removed remnant adapter sequences from the PacBio HiFi reads using HiFiAdapterFilt (Sim et al. 2022) and generated the initial dual or partially phased diploid assembly (http://lh3.github.io/2021/10/10/introducing-dual-assembly) with the filtered PacBio reads and the Omni-C data using HiFiasm (Cheng et al. 2022). We tagged output haplotype 1 as the primary assembly, and output haplotype 2 as the alternate assembly. Next, we scaffolded both assemblies using Omni-C data with SALSA (Ghurye et al. 2017, 2019).

We generated the Omni-C contact maps for both assemblies by aligning the Omni-C data against the corresponding assemblies with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools (Goloborodko et al. 2018). We generated a multiresolution Omni-C matrix with cooler (Abdenur and Mirny 2020) and balanced it with hicExplorer (Wolff et al. 2018). We used HiGlass (Kerpedjiev et al. 2018) and the PretextSuite

---

**Fig. 1.** *Anthopleura sola* polyp in sandy habitat (image credit: B. Cornwell).
To visualize the contact map. We analyzed the contact maps for major misassemblies, cutting scaffolds at the joins (gaps) where misassemblies were identified. No further joins were made after this step. Using the PacBio HiFi reads and YAGCloser, we closed some of the remaining gaps generated during scaffolding. We then checked for contamination using the BlobToolKit Framework (Challis et al. 2020). Finally, we trimmed remnants of sequence raptors and mitochondrial contamination identified during the contamination screening performed by NCBI.

Mitochondrial genome assembly
We assembled the mitochondrial genome of *A. sola* from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (https://github.com/marceluliano/MitoHiFi; Allio et al. 2020). The mitochondrial sequence of *Anthopleura midori* (NCBI:NC_030274.1) was used as the starting sequence. After completion of the nuclear genome, we searched for matches of the resulting mitochondrial assembly sequence in the nuclear genome assembly using BLAST+ (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.

**Table 1. Assembly pipeline and software used.**

| Assembly                  | Software and options                          | Version         |
|---------------------------|-----------------------------------------------|-----------------|
| Filtering PacBio HiFi adapters | HiFiAdapterFilt                              | Commit 64d1c7b |
| K-mer counting            | Meryl (k = 21)                                | 1               |
| Estimation of genome size and heterozygosity | GenomeScope                                   | 2               |
| De novo assembly (contiging) | HiFiasm (Hi-C Mode, –primary, output p_ctg.hap1, p_ctg.hap2) | 0.16.1-r375    |
| Scaffolding               |                                               |                 |
| Omni-C scaffolding        | SALSA (-DNASE, -i 20, -p yes)                 | 2               |
| Gap closing               | YAGCloser (-mins 2 -f 20 -mcc 2 -prt 0.25 -eft 0.2 -pld 0.2) | Commit 0e34c3b |
| Omni-C contact map generation |                                               |                 |
| Short-read alignment      | BWA-MEM (-5SP)                                | 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 (hicCorrectmatrix correct --filterThreshold -2.4) | 3.6             |
| Contact map visualization | HiGlass                                       | 2.1.11          |
|                                               | PretextMap                                    | 0.1.4           |
|                                               | PretextView                                   | 0.1.5           |
|                                               | PretextSnapshot                               | 0.0.3           |
| Organelle assembly        |                                               |                 |
| Mitogenome assembly       | MitoHiFi (-r,-p 50,-o 1)                      | 2 commit c06ed3e|
| Genome quality assessment |                                               |                 |
| Basic assembly metrics    | QUAST (--est-ref-size)                        | 5.0.2           |
| Assembly completeness     | BUSCO (-m geno,-l metazoa)                    | 5.0.0           |
|                                               | Merqury                                       | 2020-01-29      |
| Contamination screening   |                                               |                 |
| Local alignment tool      | BLAST+                                        | 2.1             |
| General contamination screening | BlobToolKit                                   | 2.3.3           |

Software citations are listed in the text.

‘Options detailed for nondefault parameters

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marb1/meryl). The k-mer database was then used in GenomeScope 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 (Gurevich et al. 2013). To evaluate genome quality and completeness we used BUSCO (Manni et al. 2021) with the metazoa 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 (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).
Measurements of the size of the phased blocks are based on the size of the contigs generated by HiFiasm on HiC mode. We follow the quality metric nomenclature established by Rhie et al. (2021), with the genome quality code $x\cdot y\cdot P\cdot Q\cdot C$, where $x = \log_{10}\text{contig NG50}$; $y = \log_{10}\text{scaffold NG50}$; $P = \log_{10}\text{phased block NG50}$; $Q = \text{Phred base accuracy QV (quality value)}$; $C = \%$ genome represented by the first “n” scaffolds, where $n = 28$. As there is no karyotype information available for A. sola, and karyotype of ancestral taxa varies (Genome On A Tree; https://goat.genomehubs.org/—search: “tax_name(anthopleura sola)” we are using an estimated “n” (number of chromosomes) based on scaffold size, and visual inspection of the contact maps. For consistency with nomenclature and literature we are keeping the quality code as is. Quality metrics for the notation were calculated on the primary assembly.

**Results**

The Omni-C and PacBio HiFi sequencing libraries generated 61.1 million read pairs and 3.02 million reads, respectively. The latter yielded 131.8-fold coverage (N50 read length 15,671 bp; minimum read length 47 bp; mean read length 15,656 bp; and maximum read length 53,294 bp) based on the GenomeScope 2.0 genome size estimation of 240 Mb. Based on PacBio HiFi reads, we estimated 0.274% sequencing error rate and 2.7% heterozygosity rate. The k-mer spectrum based on PacBio HiFi reads shows a bimodal distribution with 2
major peaks at ~62- and ~133-fold coverage, where peaks correspond to homozygous and heterozygous states of a diploid species, respectively (Fig. 2A). The distribution presented in this k-mer spectrum supports that of a high heterozygosity profile.

The final assembly (jaAntSola1) consists of 2 pseudo haplotypes, primary and alternate. The primary assembly has a total length of 288,960,535 bp with contig and scaffold N50 of 2,720,395 and 10,852,815 bp, respectively. We also generated an alternate assembly with similar results in total length (299,680,816 bp), but lower contig and scaffold N50s of 2,204,677 and 8,187,693 bp. The total number of contigs in the primary assembly is 368, which are assembled into 270 scaffolds; the alternate assembly contains about twice the number of assembled contigs and scaffolds: 666 and 556, respectively. The average GC content for both the primary and alternate assemblies is 38%. For the primary assembly, the longest contig was 6,637,320 while the largest scaffold was 23,766,007 bp. The largest contig in the alternative assembly is similarly sized (6,654,740), although the longest scaffold is ca. 4 Mb longer (27,475,960). On average there are 339 gaps per Gb in the primary assembly. The BUSCO scores revealed a high level of completeness for both the primary and alternative assemblies (95.60%), while duplicated (0.70%),
Sequencing Laboratory at UC Berkeley, supported by NIH sequencing platforms at the Vincent J. Coates Genomics UC Davis Genome Center, supported by National Institutes of Health. PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the UC Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high-quality sequence data.

Acknowledgments
The authors thank the staff at the UC Davis DNA Technologies and Expression Analysis Cores and the UC Santa Cruz Paleogenomics Laboratory for their contribution to this project. We also acknowledge the support 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 2020.

References
Abdenur N, Mirny LA. Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics. 2020;36(1):311–316.
Allio R, Schomaker-Bastos A, Romiguer J, Prodocimi F, Nahholz B, Delsuc F. Mitofinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Mol Ecol Resour. 2020;20(4):892–905.
Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinf. 2009;10:421. doi:10.1186/1471-2105-10-421.
Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit interactive quality assessment of genome assemblies. G3 Genes|Genomes|Genetics. 2020;10:1361–1374. doi:10.1534/g3.119.400908.
Cheng H, Jarvis ED, Fedrigo O, Koepli K-P, Urban L, Gemmell NJ, et al. Haplotype-resolved assembly of diploid individuals without parental data, Nat Biotechnol 2022;40:1332–1335. doi:10.1038/s41587-022-01261-x.
Cornwell BH. Gene flow in the anemone Anthopleura elegantissima limits signatures of local adaptation across an extensive geographic range. Mol Ecol. 2020;2020(29):2550–2566.
Cornwell BH, Hernández L. Genetic structure in the endosymbiont Breviolum muscatinei is correlated with geographical location, environment and host species. Proc R Soc B. 2021;288:20202896.
Denny M, Gaines S. Encyclopedia of tidepools and rocky shores. Berkeley (CA): University of California Press; 2007.
Dimond JL, Nguyen N, Roberts SB. DNA methylation profiling of a cnidarian-algal symbiosis using nanopore sequencing. G3 Genes|Genomes|Genetics. 2021;11(7):jkaa148.
Fuller ZL, Mocellin VJL, Morris LA, Cantin N, Shepherd J, Sarre L, Peng J, Liao Y, Pickrell J, Andolfatto P. Population genetics of the coral Acropora millepora: toward genomic prediction of bleaching. Science. 2020;369(6501):eaab4674.
Ghurye J, Pop M, Koren S, Bickhart D, Chin C-S. Scaffoldining of long read assemblies using long range contact information. BMC Genomics. 2017;18(1):527. doi:10.1186/s12864-017-3879-z.
Ghurye J, Rhie A, Walenz BP, Schmitt A, Selvaraj S, Pop M, Philippy AM, Koren, S. Integrating Hi-C links with assembly graphs for chromosome-scale assembly. PLoS Comput Biol. 2019;e1007273. doi:10.1371/journal.pcbi.1007273.
Goloborodko A, Abdennur N, Venes S, hbbrandao, gfudenberg. mirnylab/pairtools: v0.2.0. 2018. doi:10.5281/zenodo.1490831.
Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29(8):1072–1075.
Kerpedjiev P, Abdennur N, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, et al. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol*. 2018;19(1):125. doi:10.1186/s13059-018-1486-1.
Korlach J, Gedman G, Kingan SB, Chin C-S, Howard JT, Audet J-N, Jarvis ED. De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. *GigaScience*. 2017;6(10):1–16.
Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*, 2013;arXiv:1303.3997 , preprint: not peer reviewed. doi:10.48550/arXiv.1303.3997.
López-Nandam EH, Albright R, Hanson EA, Sheets EA, Palumbi SR. Mutations in coral soma and sperm imply lifelong stem cell renewal and cell lineage selection, *bioRxiv*, 2021, 453148, preprint: not peer reviewed. doi:10.1101/2021.07.20.453148.
Manni M, Berkeley MR, Seppey M, Siafo FA, Zdobnov EM. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes, *Mol Biol Evol*, 2021;38:4647–4654. doi:10.1093/molbev/msab199.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. 2020;11(1):1432. doi:10.1038/s41467-020-14998-3.
Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammashan A, Kim J, et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature*. 2021;592:737–746. doi:10.1038/s41586-021-03451-0.
Rhie A, Walenz BP, Koren S, Philippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21(1):245. doi:10.1186/s13059-020-02134-9.
Sanford E, Sones JL, García-Reyes M, et al. Widespread shifts in the coastal biota of northern California during the 2014–2016 marine heatwaves. *Sci Rep*. 2019;9:4216. doi:10.1038/s41598-019-40784-3.
Sim SB, Corpuz RL, Simmonds TJ, Geib SM. 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*. 2022;23:157. doi:10.1186/s12864-022-08375-1.
Wolff J, Bhardwaj V, Nothjunge S, Richard G, Renschler G, Gilshbach R, Manke T, Backofen R, Ramirez F, Grüning BA. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Res*. 2018;46:W11–W16. doi:10.1093/nar/gky504.