A draft reference genome of the red abalone, *Haliotis rufescens*, for conservation genomics

Joanna S. Griffiths¹, Ruta M. Sahasrabudhe², Mohan P.A. Marimuthu², Oanh H. Nguyen², Eric Beraut³, Merly Escalona⁴, Andrew Whitehead¹

¹: Department of Environmental Toxicology, University of California Davis, Davis, CA, 95616
²: DNA Technologies and Expression Analysis Core Laboratory, Genome Center, University of California-Davis, CA 95616, USA
³: Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064, USA
⁴: Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, 95691, USA

*Corresponding author email:* jsgriffiths@ucdavis.edu

© The American Genetic Association. 2022.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Abstract

Red abalone, *Haliotis rufencens*, are herbivorous marine gastropods that primarily feed on kelp. They are the largest and longest-lived of abalone species with a range distribution in North America from central Oregon, USA, to Baja California, MEX. Recently, red abalone have been in decline as a consequence of over-harvesting, disease, and climate change, resulting in the closure of the commercial fishery in the 1990s and the recreational fishery in 2018. Protecting this ecologically and economically important species requires an understanding of their current population dynamics and connectivity. Here, we present a new red abalone reference genome as part of the California Conservation Genomics project (CCGP). Following the CCGP genome strategy, we used Pacific Biosciences HiFi long-reads and Dovetail Omni-C data to generate a scaffold-level assembly. The assembly comprises 616 scaffolds for a total size of 1.3 Gb, a scaffold N50 of 45.7 Mb, and a BUSCO complete score of 97.3%. This genome represents a significant improvement over a previous assembly and will serve as a powerful tool for investigating seascape genomic diversity, local adaptation to temperature and ocean acidification, and informing management strategies.

Keywords

California Conservation Genomics Project, CCGP, red abalone, conservation genomics
Introduction

Red abalone, *Haliotis rufescens*, are an invaluable resource that have been used by Native coastal Californians (Rick, Braje, and Erlandson 2019) and have been commercially and recreationally harvested along the West coast of the United States. Their ability to grow quickly and reach large sizes has supported a fledgling aquaculture industry (Cook 2016). Red abalone have the broadest geographic range out of all seven *Haliotis* species native to California, extending from central Oregon, USA, to Baja California, MEX (Cox 1962). Recently, however, red abalone have been impacted by overharvesting, climate change, and disease, which has resulted in population declines and the closing of the recreational fishery in 2018 (Rogers-Bennett and Catton 2019). Warming ocean temperatures have increased the prevalence of the bacterial disease “withering abalone syndrome”, further contributing to abalone declines. Their main food source, bull kelp, has also declined from sea urchin population explosions and warming temperatures (Rogers-Bennett and Catton 2019). Robust genomic resources for this species would provide an important foundation for genomics-enabled research that may enhance aquaculture production and sustainability, and improve conservation efforts in the wild. For example, wild red abalone populations show varying sensitivities to ocean acidification exposure (Swezey et al. 2020), suggesting that red abalone harbors genetic variation that may provide resilience to changing oceans. Identification of adaptively important genetic variation could be useful to help guide sustainable aquaculture and wild conservation.

Here, we report on the genome assembly of red abalone as part of the California Conservation Genomics Project (CCGP), the goal of which is to use a “community genomics” approach to describe patterns across 230 species native to California (Shaffer et al. 2022). Our goal was to generate a reference genome that is an improvement on a previous assembly (Masonbrink et al. 2019), and of quality comparable to other species included in the CCGP. Using the newly generated reference genomes and re-sequenced individuals from across the state, the CCGP are developing tools to identify important hotspots of genetic diversity for multiple species and
providing a framework for informed conservation decisions and management plans. The CCGP have successfully assembled the genomes for a few species, including the black abalone, northwestern pond turtle, and Big Berry Manzanita (Huang et al. 2022; Todd et al. 2022; Orland et al. 2022). With support from CCGP, we generated a scaffold-level assembly using a hybrid de novo approach that combines Hi-C chromatin-proximity and PacBio HiFi long-read sequencing data.

**Methods**

**Biological Materials:**

Adult red abalone from Van Damme State Park, CA (39.269 N, 123.798 W), were collected in 2016 and housed at Bodega Marine Laboratory (University of California Davis). Foot and epipodial clippings (100-200mg) from two abalone individuals were collected in July of 2020 and 2021 for DNA extraction and generation of the reference genome.

**Nucleic acid library preparation and sequencing**

**High molecular weight DNA extraction**

High molecular weight (HMW) genomic DNA (gDNA) was extracted from 36mg of the epipodial tissue collected in 2021 using the Nanobind Tissue Big DNA kit (Pacific Biosciences - PacBio, CA) following the manufacturer’s instructions with the following minor modifications. We performed an additional wash of tissue homogenate with the CT buffer and pelleted it by centrifuging at 18,000g (4°C for 5 minutes) to remove residual buffer before proceeding with the lysis step. The extracted HMW DNA was further purified using the phenol-chloroform extraction method (PacBio). We assessed DNA purity using absorbance ratios (260/280 = 1.83 and 260/230 = 2.09) on a NanoDrop ND-1000 spectrophotometer. The DNA yield (148 ng/μl; 11.1μg total) was quantified
using QuantiFluor ONE dsDNA Dye assay (Promega, WI). We determined the size distribution of the HMW DNA using the Femto Pulse system (Agilent, CA) and found that 83% of the fragments were 100 kb or more.

**HiFi library preparation and sequencing**

The HiFi SMRTbell library was constructed using the SMRTbell Express Template Prep Kit v2.0 (PacBio, Cat. #100-938-900) according to the manufacturer's instructions. HMW gDNA was sheared to a target DNA size distribution between 15 kb – 18 kb. The sheared gDNA was concentrated using 0.45X of AMPure PB beads (PacBio, Cat. #100-265-900) for the removal of single-strand overhangs at 37°C for 15 minutes, followed by further enzymatic steps of DNA damage repair at 37°C for 30 minutes, end repair and A-tailing at 20°C for 10 minutes and 65°C for 30 minutes, and ligation of overhang adapter v3 at 20°C for 60. The SMRTbell library was purified and concentrated with 1X Ampure PB beads (PacBio, Cat. #100-265-900) for nuclease treatment at 37°C for 30 minutes and size selection using the BluePippin/PippinHT system (Sage Science, MA; Cat #BLF7510/HPE7510) to collect fragments greater than 7.9 kb. The 15 – 20 kb average HiFi SMRTbell library was sequenced at University of California Davis DNA Technologies Core (Davis, CA) using two 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-hour movies each on a PacBio Sequel II sequencer.

**Omni-C library preparation and sequencing**

The Omni-C library was prepared using the Dovetail™ Omni-C™ Kit (Dovetail Genomics, CA) according to the manufacturer’s protocol with slight modifications. First, the foot tissue collected in 2020 was thoroughly ground with a mortar and pestle while cooled with liquid nitrogen. Subsequently, chromatin was fixed in place in the nucleus. The suspended chromatin solution was then passed through 100 μm and 40 μm cell strainers to remove large debris. Fixed chromatin was
digested under various conditions of DNase I until a suitable fragment length distribution of DNA molecules was obtained. Chromatin ends were repaired and ligated 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. Purified DNA was treated to remove biotin that was not internal to ligated fragments. An NGS library was generated using an NEB Ultra II DNA Library Prep kit (New England Biolabs, MA) with an Illumina compatible y-adaptor. Biotin-containing fragments were then captured using streptavidin beads. The post-capture product was split into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The libraries were sequenced at Vincent J. Coates Genomics Sequencing Lab (Berkeley, CA) on an Illumina NovaSeq platform (Illumina, CA) to generate approximately 100 million 2 x 150 bp read pairs per GB of genome size.

**Genome Assembly**

**Nuclear genome assembly**

We assembled the red abalone genome following the CCGP assembly protocol Version 4.0, as outlined in Table 1. As with other CCGP assemblies, our goal is to produce a high quality and highly contiguous assembly using PacBio HiFi reads and Omni-C data while minimizing manual curation. Briefly, we removed remnant adapter sequences from the PacBio HiFi dataset using HiFiAdapterFilt (Sim 2021) and obtained the initial dual or partially phased diploid assembly (http://lh3.github.io/2021/10/10/introducing-dual-assembly) with HiFiasm using the filtered HiFi reads and the Omni-C reads (Cheng et al. 2021). We tagged output haplotype 1 as the primary assembly, and output haplotype 2 as the alternate assembly. We scaffolded both assemblies using the Omni-C data with SALSA(Ghurye et al. 2017; Ghurye et al. 2019).
We generated Omni-C contact maps for both assemblies by aligning the Omni-C data against the corresponding assembly with BWA-MEM (Li 2013), identified ligation junctions, and generated Omni-C pairs using pairtools (Goloborodko et al. 2018). We generated a multi-resolution Omni-C matrix with cooler (Abdennur and Mirny 2020) and balanced it with hicExplorer [Version 3.6] (Ramírez et al. 2018). We used HiGlass (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. We checked the contact maps for major mis-assemblies. If found, we cut the assemblies at the gaps where misassemblies were found. No further joins were made after this step. Using the PacBio HiFi reads and YAGClosrer (https://github.com/merlyescalona/yagclosrer), 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 adaptors and mitochondrial contamination. The genome was automatically run through NCBI’s Eukaryotic Genome Annotation Pipeline using RNA-Seq reads from (Masonbrink et al. 2019).

Mitochondrial genome assembly

We assembled the mitochondrial genome of the red abalone from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (https://github.com/marcelauliano/MitoHiFi; Allio et al. 2020). The mitochondrial sequence of *Haliotis ovina* (NCBI: MZ147805.1) 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+ (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.
**Genome size estimation and quality assessment**

We generated k-mer counts (k=21) from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). The k-mer database was then used in GenomeScope2.0 (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 is based on the size of the contigs generated by HiFiasm on HiC mode (initial diploid assembly). Following data availability and quality metrics established by Rhie et al. (2021), we will use the derived genome quality notation x:y:P:Q:C, where, x = log10(contig NG50); y = log10[scaffold NG50]; P = log10 [phased block NG50]; Q = Phred base accuracy QV (quality value); C = % genome represented by the first ‘n’ scaffolds, following a known karyotype of 2n=36 (Gallardo-Escarate et al. 2004). Quality metrics for the notation were calculated on the primary assembly.

**Results**

The Omni-C and PacBio HiFi sequencing libraries generated 52.2 million read pairs and 4.3 million reads respectively. The latter yielded 50.8-fold coverage (N50 read length 15,755 bp; minimum read length 242 bp; mean read length 15,032 bp; maximum read length 58,582 bp) based on the Genomescope2.0 genome size estimation of 1.28 Gb. Based on PacBio HiFi reads, we estimated 0.139 % sequencing error rate and 1.37 % heterozygosity rate. The k-mer spectrum based on PacBio HiFi reads show (Figure 2A) a bimodal distribution with two major peaks at ~24- and ~49-fold coverage, where peaks correspond to homozygous and heterozygous states of a diploid species. The distribution presented in this k-mer spectrum supports that of a high heterozygosity profile.
The final assembly (xgHalRufe1) consists of two pseudo haplotypes, primary and alternate, where both genome sizes are similar to the estimated value from Genomescope2.0 (Figure 2A). The primary assembly consists of 616 scaffolds spanning 1.33 Gb with contig N50 of 8.8 Mb, scaffold N50 of 45.7 Mb, largest contig of 38.6 Mb and largest scaffold of 94.2 Mb. On the other hand, the alternate assembly consists of 494 scaffolds, spanning 1.37 Gb with contig N50 of 7.9 Mb, scaffold N50 of 44.1 Mb, largest contig of 35.6 Mb and largest scaffold of 78.7 Mb. Assembly statistics are reported in tabular and graphical forms in Table 2, and Figure 2B for the primary assembly (See Supplementary Figure 1 for the alternate assembly).

We identified a total of 16 misassemblies generated during the scaffolding step, 7 on the primary, and 9 on the alternate assembly, and broke the assemblies at the corresponding joins. We were able to close a total of 24 gaps, 9 on the primary assembly and 15 on the alternate assembly. We filtered 2 contigs from the primary assembly corresponding to contamination matches to a nematode and a brachiopod, and 1 contig from the alternate corresponding to an arthropod contaminant. Finally, we filtered out a single contig from the alternate assembly corresponding to mitochondrial contamination. The primary assembly has a BUSCO completeness score of 97.3% using the metazoa gene set, a per base quality (QV) of 64.9, a kmer completeness of 81.3, and a frameshift indel QV of 50.4. The alternate assembly has a BUSCO completeness score of 97.4% using the aves gene set, a per base quality (QV) of 65.6, a kmer completeness of 82.7, and a frameshift indel QV of 50.1. The Omni-C contact maps shows that both assemblies are highly contiguous (Figure 2C & 2D). We have deposited scaffolds corresponding to both primary and alternate haplotypes on NCBI (See Table 2 and Data availability for details).

The mitochondrial genome assembled with MitoHiFi has a genome size of 17,141 bp. The base composition of the mitochondrial assembly is A=25.44%, C=13.62%, G=25.45%, T=35.49% and consists of 22 unique transfer RNAs and 13 protein coding genes.
Discussion

The previously assembled genome for red abalone, using PacBio long reads and Illumina short reads, comprised 8,371 scaffolds with a BUSCO complete score of 95.1%, suggesting that the assembly is of high quality (Masonbrink et al. 2019). To be consistent with the CCGP data analysis pipeline and to ensure that our dataset is comparable with other species from CCGP using the “community genomics” approach, we generated a new reference genome using a hybrid assembly that combines PacBio long read data with Omni-C chromatin conformation data used for scaffolding. Compared to the previous assembly, the new reference genome is composed of a considerably smaller number of scaffolds (615) and an improved BUSCO complete score of 97.3%.

We found the genome size of *H. rufescens* to be 1.33 Gb, which is similar to the previous estimate for this species (1.49 Gb; Masonbrink et al. 2019), and to other *Haliotis* species, such as *H. rubra* (blacklip abalone 1.24 - 1.31 Gb; Gan et al. 2019), *H. laevigata* (greenlip abalone 1.76 Gb; Botwright et al. 2019), *H. cracherodii* (black abalone 1.1 Gb; Orland et al. 2022), and *H. discus hannai* (Pacific abalone 1.86 Gb; Nam et al. 2017). The GC content for our new genome was 40.9%, a value similar to the previously assembled genome by Masonbrink et al. (2019)(40.2%).

This reference genome will serve as an important resource for ongoing investigations using whole-genome resequencing efforts for populations spanning southern Oregon, USA, to Baja California, MEX, as part of the CCGP goals (Shaffer et al. 2022). We plan to address important questions that will inform conservation and management practices: 1) What is the genetic structure of red abalone populations and the degree of population connectivity? We are particularly interested in investigating the level of gene flow among northern California and southern Oregon populations where connectivity among populations is virtually unknown (Gruenthal, Acheson, and Burton 2007). 2) What is the genetic basis of adaptation to temperature and ocean acidification resilience? The red abalone’s distribution spans a broad geographic range, exposing them to a
variety of environmental gradients, which may promote local adaptation. 3) Using historical and modern samples, what is the effective population size for populations in recent decline? Given the difficulty of performing scuba transect surveys and the low density of individuals in their habitat, genomic data may be able to fill this knowledge gap. The outcomes of research studies that address these questions will provide important insights to help guide conservation, management, and sustainable aquaculture for this iconic species.
Funding

This work was supported by the California Conservation Genomics Project, with funding provided to the University of California by the State of California, State Budget Act of 2019 [UC Award ID RSI-19-690224 to AW].

Acknowledgements

We thank Blythe Marshman and Dr. Sara Boles for collecting epipodal tissue used to generate the reference genome. PacBio Sequel II library prep and sequencing was carried out at the DNA Technologies and Expression Analysis Cores at the University of California 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 University of California Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank the staff at the University of California Davis DNA Technologies and Expression Analysis Cores and the University of California Santa Cruz Paleogenomics Laboratory for their diligence and dedication to generating high quality sequence data. We thank the CCGP leadership team (Brad Schaffer, Victoria Sork, Erin Toffelmier) for securing funding and for project organization and coordination.

Data Availability

Data generated for this study are available under NCBI BioProject PRJNA777175. Raw sequencing data for VD_foot and VD_epi2 (NCBI BioSample SAMN26275698, SAMN26275699) are deposited in the NCBI Short Read Archive (SRA) under SRX15312148 for the PacBio HiFi sequencing data and SRX15312149, SRX15312150 for the Omni-C sequencing data. GeneBank accessions for both primary and alternate assemblies are GCA_023055435.1 and GCA_023055495.1; and genome
sequences JALGQA0000000000 and JALGQB0000000000. The GenBank accession for the mitochondrial genome is JALGQA010000616.1. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository:

www.github.com/ccgproject/ccgp_assembly. RNA-Seq read data used to annotate the genome are located in the NCBI short read archive (BioProject accession: PRJNA488641).
References

Abdennur, N., & Mirny, L. A. (2020). Cooler: scalable storage for Hi-C data and other genomically labeled arrays. Bioinformatics, 36(1), 311–316. doi: 10.1093/bioinformatics/btz540

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(4), 892–905. doi: 10.1111/1755-0998.13160

Botwright, Natasha A., Min Zhao, Tianfang Wang, Sean McWilliam, Michelle L. Colgrave, Ondrej Hlinka, Sean Li, et al. 2019. “Greenlip Abalone Genome and Protein Analysis Provides Insights into Maturation and Spawning.” G3 9 (10): 3067–78.

Cox, Keith W. 1962. “Fish Bulletin No. 118. California Abalones, Family Haliotidae,” May. https://escholarship.org/content/qt5c46p19z/qt5c46p19z.pdf?t=krnkos.

Gan, Han Ming, Mun Hua Tan, Christopher M. Austin, Craig D. H. Sherman, Yen Ting Wong, Jan Strugnell, Mark Gervis, Luke McPherson, and Adam D. Miller. 2019. “Best Foot Forward: Nanopore Long Reads, Hybrid Meta-Assembly, and Haplotig Purging Optimizes the First Genome Assembly for the Southern Hemisphere Blacklip Abalone.” Frontiers in Genetics 10: 889.

Gruenthal, K. M., L. K. Acheson, and R. S. Burton. 2007. “Genetic Structure of Natural Populations of California Red Abalone (Haliotis Rufescens) Using Multiple Genetic Markers.” Marine Biology 152 (6): 1237–48.

Huang, Yi, Merly Escalona, Glen Morrison, Mohan P. A. Marimuthu, Oanh Nguyen, Erin Toffelmier, H. Bradley Shaffer, and Amy Litt. 2022. “Reference Genome Assembly of the Big Berry Manzanita (Arctostaphylos Glauca).” The Journal of Heredity 113 (2): 188–96.

Masonbrink, Rick E., Catherine M. Purcell, Sara E. Boles, Andrew Whitehead, John R. Hyde, Arun S. Seetharam, and Andrew J. Severin. 2019. “An Annotated Genome for Haliotis Rufescens (Red Abalone) and Resequence Green, Pink, Pinto, Black, and White Abalone Species.” Genome Biology and Evolution 11 (2): 431–38.

Nam, Bo-Hye, Woori Kwak, Young-Ok Kim, Dong-Gyun Kim, Hee Jeong Kong, Woo-Jin Kim, Jeong-Ha Kang, et al. 2017. “Genome Sequence of Pacific Abalone (Haliotis
Orland, Chloé, Merly Escalona, Ruta Sahasrabudhe, Mohan P. A. Marimuthu, Oanh Nguyen, Eric Beraut, Blythe Marshman, James Moore, Peter Raimondi, and Beth Shapiro. 2022. “A Draft Reference Genome Assembly of the Critically Endangered Black Abalone, Haliotis Cracherodii.” The Journal of Heredity, May, esac024.

Rick, Torben C., Todd J. Braje, and Jon M. Erlandson. 2019. “Early Red Abalone Shell Middens, Human Subsistence, and Environmental Change on California’s Northern Channel Islands.” Journal of Ethnobiology 39 (2): 204.

Rogers-Bennett, L., and C. A. Catton. 2019. “Marine Heat Wave and Multiple Stressors Tip Bull Kelp Forest to Sea Urchin Barrens.” Scientific Reports 9 (1): 15050.

Shaffer, H. Bradley, Erin Toffelmier, Russ B. Corbett-Detig, Merly Escalona, Bjorn Erickson, Peggy Fiedler, Mark Gold, et al. 2022. “Landscape Genomics to Enable Conservation Actions: The California Conservation Genomics Project.” The Journal of Heredity, April, esac020.

Swezey, Daniel S., Sara E. Boles, Kristin M. Aquilino, Haley K. Stott, Doug Bush, Andrew Whitehead, Laura Rogers-Bennett, Tessa M. Hill, and Eric Sanford. 2020. “Evolved Differences in Energy Metabolism and Growth Dictate the Impacts of Ocean Acidification on Abalone Aquaculture.” Proceedings of the National Academy of Sciences of the United States of America 117 (42): 26513–19.

Todd, Brian D., Thomas S. Jenkinson, Merly Escalona, Eric Beraut, Oanh Nguyen, Ruta Sahasrabudhe, Peter A. Scott, Erin Toffelmier, Ian J. Wang, and H. B. Shaffer. 2022. “Reference Genome of the Northwestern Pond Turtle, Actinemys Marmorata.” The Journal of Heredity, June, esac021.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.

BLAST+: architecture and applications. BMC Bioinformatics. 10:421.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. 2020. BlobToolKit – Interactive Quality
Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, et al. Robust haplotype-resolved assembly of diploid individuals without parental data. arXiv [q-bio.GN]. 2021. Available: http://arxiv.org/abs/2109.04785

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), 527. doi: 10.1186/s12864-017-3879-z

Ghurye, J., Rhie, A., Walenz, B. P., Schmitt, A., Selvaraj, S., Pop, M., ... Koren, S. (2018). Integrating Hi-C links with assembly graphs for chromosome-scale assembly (p.e1007273). doi: 10.1371/journal.pcbi.1007273

Goloborodko, A., Abdennur, N., Venev, S., hbbrandao, & gfudenberg. (2018). mirnylab/pairtools: v0.2.0. doi: 10.5281/zenodo.1490831

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072–1075. doi: 10.1093/bioinformatics/btt086

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., ... Gehlenborg, N. (2018). HiGlass: web-based visual exploration and analysis of genome interaction maps. Genome Biology, 19(1), 125. doi: 10.1186/s13059-018-1486-1

Korlach, J., Gedman, G., Kingan, S. B., Chin, C.-S., Howard, J. T., Audet, J.-N., ... Jarvis, E. D. (2017). De novo PacBio long-read and phased avian genome assemblies correct and add to reference genes generated with intermediate and short reads. GigaScience,
Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

arXiv. Retrieved from http://arxiv.org/abs/1303.3997

Manni M., Berkeley M.R., Seppey M., Simao F.A., Zdobnov E.M. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. arXiv:2106.11799 [q-bio] [Internet]. Available from: http://arxiv.org/abs/2106.11799

Ranallo-Benavidez, T. R., Jaron, K. S., & Schatz, M. C. (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nature Communications, 11(1), 1432. doi: 10.1038/s41467-020-14998-3

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biology, 21(1), 245. doi: 10.1186/s13059-020-02134-9

Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, Uliano-Silva M, Chow W, Fungtammasan A, Kim J, et al. (2021). Towards complete and error-free genome assemblies of all vertebrate species. Nature 592:737–746. doi: 10.1038/s41586-021-03451-0

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: 157. doi:10.1186/s12864-022-08375-1
Figure and Table Captions

Figure 1. A: Adult red abalone, *Haliotis rufescens*. Photo taken by Jackson Gross. B: Red abalone larvae 74 days post-fertilization at 1.6x magnification. Photo taken by Sara Boles.

Figure 2. Visual overview of genome assembly metrics. A) K-mer spectrum 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 high frequency correspond to differences between haplotypes, whereas the higher coverage and slightly lower 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 *Haliotis rufescens* primary assembly (xgHalRufe1). 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 versus AT content at 0.1% intervals (Challis et al. 2020). C- D) Omni-C Contact maps for the primary (2C) and alternate (2D) genome assembly generated with PretextSnapshot. Hi-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 of such regions.
Table 1: Assembly Pipeline and Software Used. Software citations are listed in the text.

Table 2: Sequencing and assembly statistics, and accession numbers.
Table 1: Assembly Pipeline and Software Used. Software citations are listed in the text. § Options detailed for non-default runs.

| Assembly                                      | Software and options § | Version |
|-----------------------------------------------|-------------------------|---------|
| Filtering PacBio HiFi adapters                | HiFiAdapterFilter       | 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_cpg hap1, p_cpg hap2) | 0.16.1-r375 |
| **Scaffolding**                               |                         |         |
| Omni-C Scaffolding                           | SALSA (-DNASE, -i 20, -p yes) | 2       |
| Gap closing                                   | YACCloser (-min 2 -f 20 -moc 2 -p1 0.25 -e 0.2 -p2 0.2) | commit 0654c3b |
| **Ommi-C Contact map generation**             |                         |         |
| Short-read alignment                         | SWA-MEM (-ESP)          | 0.7.17@1188 |
| SAM/BAM processing                           | samtools                | 1.3.1   |
| SAM/BAM filtering                            | pairtools               | 0.3.0   |
| Pairs indexing                               | paivar                  | 0.3.7   |
| Matrix generation                            | cooler                   | 0.8.10  |
| Matrix balancing                             | hicExplorer (hicCorrectmatrix correct --filterThreshold -2.4) | 3.6     |
|                                              | HiGlass                  | 2.1.11  |
|                                              | PerexMap                 | 0.1.4   |
|                                              | Perex3View               | 0.1.5   |
| Contact map visualization                    | Perex3Snapshot          | 0.0.3   |
| **Organelle assembly**                       |                         |         |
| Mitogenome assembly                          | MitohiFi (-r, -p 50, -o 1) | 2 commit c0bed3e |
| **Genome quality assessment**                 |                         |         |
| Basic assembly metrics                        | QUAST (--est-ref-size)  | 5.0.2   |
| Assembly completeness                         | BUSCO (m geno, 4 metazo) | 5.0.0   |
|                                              | Merqury                  | 1       |
| **Contamination screening**                   |                         |         |
| Local alignment tool                          | BLASTX                   | 2.10    |
| General contamination screening               | BlobToolKit             | 2.3.3   |
Table 2: Sequencing and assembly statistics, and accession numbers.

| Bio Projects & Vouchers | Primary | Alternate |
|-------------------------|---------|-----------|
| CCGBP NCEB Bioproject   | PRJNA720569 |           |
| Genera NCEB Bioproject  | PRJNA76538  |           |
| Species NCEB Bioproject | PRJNA777175 |           |
| NCEB BioSample          | SAMN6275698, SAMN6275699 |       |
| Specimen identification | Individual from Van Damme State park (USA-California) |       |
| NCEB Genome accessions  |         |           |
| Assembly accession      | JALGQA00000000 | JALGQB00000000 |
| Genome sequences         | GCA_023055435.1 | GCA_023055495.1 |

| Genome Sequence | PacBio HiFi reads | Assembly identifier (Quality code *) |
|-----------------|-----------------|-------------------------------------|
| Run             | 1 PACBIO_SMRT (Sequel II) run  | 52.3M reads, 14.7G bases, 5GB |
| Accession       | SRX15312149     | xgHalRufe1(6.7.Pe.Q99.C71) |

| Genome Assembly Quality Metrics | Primary | Alternate |
|--------------------------------|---------|-----------|
| HiFi Read coverage †           | 50.6X   |           |

| Assembly completeness (metazoan) n=914 | C     | D     | F     | M     |
|---------------------------------------|-------|-------|-------|-------|
| BUSCO completeness                     | 97.30%| 96.90%| 0.49% | 1.70% | 1.00% |
| Genomes                               | 97.40%| 96.20%| 1.20% | 1.50% | 1.10% |

* Assembly quality code x.y.P.Q.C derived notation, from (Rhie et al. 2021). x = log10[contig NG50]; y = log10[scaffold NG50]; P = log10 [phased block NG50]; Q = Phred base accuracy QV (Quality value); C = % genome represented by the first ‘n’ scaffolds, following a known karyotype of 2n=36 (Gallardo-Escarate et al. 2004).

‡ Read coverage and NGx statistics have been calculated based on the estimated genome size of 1.28 Gb.

§ P(imentary) and A(Alternate) assembly values.
