Reference genome for the California ribbed mussel, *Mytilus californianus*, an ecosystem engineer

Lisa X. Paggeot[1,*], Melissa B. DeBiasse[1], Merly Escalona[2], Colin Fairbairn[3], Mohan P. A. Marimuthu[4], Oanh Nguyen[4], Ruta Sahasrabudhe[4], Michael N Dawson[1]

[1] Life & Environmental Sciences, University of California, Merced, CA 95343.

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

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

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

ORCID: ME: 0000-0003-0213-4777, MPAM: 0000-0001-6121- 3286

*Corresponding author email: lpaggeot@ucmerced.edu

Abstract

The California ribbed mussel, *Mytilus californianus*, is an ecosystem engineer crucial for the survival of many marine species inhabiting the intertidal zone of California. Here, we describe the first reference genome for *M. californianus* and compare it to previously published genomes from three other *Mytilus* species: *M. edulis, M. coruscus,* and *M. galloprovincialis*. The *M. californianus* reference genome is 1.65 Gb in length, with N50 sequence length of 118Mb, and an estimated 86.0% complete single copy genes. Compared to the other three *Mytilus* species, the *M. californianus* genome assembly is the longest, has the highest N50 value, and the highest percentage complete single copy genes. This high-quality genome assembly provides a foundation for population genetic analyses that will give insight into future conservation work along the coast of California.

Keywords

Bivalvia, ecosystem engineer, California Conservation Genomics Project (CCGP), marine invertebrate, mussel, rocky intertidal

Introduction

Ecosystem engineers modify the physical environment in a way that changes available habitats (Jones et al. 1994). Their modifications can lead to alteration, expansion, or formation of novel habitats and promote the success of taxa in their vicinity (Dayton 1972; Bruno et al. 2003). The California ribbed mussel, *Mytilus californianus*, an intertidal species distributed along the west coast of North America from the Aleutian Islands to Isla Socorro, Mexico (Figure 1a; Soot-Ryen 1955), is an ecosystem engineer known for its role in the origin of the Keystone Species concept (Figure 1b; Paine 1966). Ribbed mussels form large compact beds, attached to underlying rock, that generate many protected interstices inhabited by other organisms (Paine 1994; Gutierrez et al. 2003). The mussel-rock attachment is by byssal threads (Waite 2017), a feature that is widespread among marine bivalves though stronger in *M. californianus* than in its congeners (Holten-Andersen et al 2009) and different in mytilids versus other bivalves (Pearce and LaBarbera 2009).

In addition to its role in important ecological functions, *Mytilus* has been a rich model system for breaking new ground in the field of marine genetics. For example, analyses of two polymorphic loci (*Lap* and *Pgi*) led Tracey et al. (1975) to suggest that homozygote excess is more apparent in juveniles than in adults due to breeding subpopulation structure within the reproductive population. Analyses of hybrid zones where distributions of *Mytilus* species overlap across the Atlantic and Pacific oceans advanced understanding of evolutionary processes in high dispersal species (Riginos and Cunningham 2004, Springer and Crespi 2006). And, exploiting the bi-parental inheritance of mtDNA in mussels, Śmietanka et al. (2010) sequenced both mitochondrial genomes of *M. trossulus* to identify that the mitogenome, which often has been treated as neutral in phylogeographic studies, likely contains multiple adaptive mutations. However, many of these previous genetic studies in *Mytilus* relied on a small number of loci, therefore limiting inference to small portions of the genome. Extending the genomic resources available would help us better understand the breadth of genomic evolution and its consequences in ecological contexts. Such genomic resources also will illuminate the genetic architecture underlying traits related to physiological tolerance and mechanisms of reproductive compatibility as well as adaptive potential of populations and species in the face of global change (Place et al. 2008; Savolainen et al. 2013; Kinoshita and Seki 2014).

Here, we present a reference genome for *M. californianus* and compare it with the previously published genomes of three other *Mytilus* species whose ranges cover the Mediterranean Sea (*M. galloprovincialis*), the coasts of China, Korea, Japan (*Mytilus coruscus*), and the western Pacific Ocean (*Mytilus edulis*). The *M. californianus* reference genome will contribute to conservation and management of regional biodiversity in California (Shaffer et al. 2022) and has the potential to provide new insights into responses to anthropogenic events and help identify consequences of local and/or regional genetic variation. In addition, the *M. californianus* reference genome will give us a better understanding of comparative genomics across multiple *Mytilus* species living around the globe.

Methods

Biological materials

One California mussel was collected from McClures Beach, Marin County, California, USA (38.1813, -122.9644) on July 22, 2020 by Michael N Dawson. The specimen was transported live to the University of California, Davis, where subsamples of dissected tissue were flash frozen in liquid nitrogen. A voucher specimen (M0D057914Y) is archived in the Dawson Lab collection at University of California, Merced.

Nucleic acid extraction, library prep, and sequencing

We extracted DNA and prepped libraries following standard methods established for marine invertebrates by the California Conservation Genomics Project (CCGP) as described in DeBiasse et al. (submitted) and available in the supplementary information. In brief, we extracted high molecular weight (HMW) DNA from 40mg of the mantle using the Nanobind Tissue Big DNA kit (Pacific BioSciences [PacBio], CA) with the following minor modifications: we performed an additional wash with the CT buffer for the tissue homogenate and pelleted it by centrifuging at 18000 x *g* (4°C for 5 minutes) to remove any residual buffer before proceeding with the lysis step. We prepared the HiFi SMRTbell library using the SMRTbell Express Template Prep Kit v2.0 (PacBio) and sequenced the 15–20Kb average HiFi SMRTbell library using three SMRT® Cell 8M Trays. We prepared the Omni-C library using the Dovetail™ Omni-C™ Kit (Dovetail Genomics, CA) according to the manufacturer's protocol with the following modifications: we optimized the digest with 2 uL of Nuclease Enzyme Mix input and the proximity ligation reaction with 500ng DNA input. We sequenced the library at the Vincent J. Coates Genomics Sequencing Laboratory at University of California, Berkeley (Berkeley, CA) on an Illumina NovaSeq platform (Illumina, CA) targeting approximately 100 million 150Bp paired end reads per gigabase of genome size.

Nuclear and mitochondrial genome assembly

We assembled the nuclear genome of *M. californianus* following the CCGP assembly protocol Version 3.0 (Table 1, Lin et al. 2022). We assembled the mitochondrial genome from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (https://github.com/marcelauliano/MitoHiFi) (Allio et al., 2020) and the *Mytilus trossulus* mitochondrial genome (NCBI:GU936626.1) as the starting reference sequence. Full assembly details are available in Lin et al. 2022.

Genome size estimation and quality assessment

We estimated genome size, heterozygosity, repeat content, sequencing error, and genome assembly completeness following standard protocols established by the CCGP and described in detail by Lin et al. (2022). We assessed assembly quality of the *M. californianus* primary and alternate genomes using BUSCO (Simão et al. 2015; Seppey et al. 2019) with the Mollusca ortholog database

(mollusca_odb10) which contains 5,295 genes. Following data availability and quality metrics established by Rhie et al. (2021), we used the derived genome quality notation x.y.Q, where, x = log10[contig NG50]; y = log10[scaffold NG50]; Q = Phred base accuracy QV (quality value); C = % genome represented by the first 'n' scaffolds, following a known karyotype of 2n = 28 (Ahmed and Sparks 1970). Quality metrics for the notation were calculated on the primary assembly.

Under the assumption that the longest scaffolds contain the majority of the genome sequence and represent the putative chromosomes, we generated a histogram of scaffold lengths for (a) the largest 20 scaffolds and (b) all scaffolds and then performed a k-means clustering in R (R Core Team 2020) to test if a drop-off in scaffold size corresponded to the number of chromosomes predicted for *Mytilus* mussels, including *M. californianus* (Ahmed and Sparks 1970, Pérez-García et al. 2014).

Comparison to previously published Mytilus *genomes*

We downloaded the complete genome sequences for the mussels *Mytilus edulis* (GCA_019925275.1), *Mytilus galloprovincialis* (GCA_900618805.1), and *Mytilus coruscus* (GCA_017311375.1) from GenBank. We calculated common metrics of assembly completeness across the three published *Mytilus* genomes and the *M. californianus* genome generated here using BUSCO [Version 5.0.0] and the Mollusca ortholog database (mollusca_odb10) as implemented in gVolante (Nishimura et al. 2019).

Results

Nucleic acid extraction, library prep, and sequencing

Extracted HMW DNA had purity 260/280 = 1.83 and 260/230 = 1.91, concentration 169 ng/µl (19.4 µg total), and good integrity with >84% of DNA fragments being 120 Kb or more. Sequencing resulted in 5.1 million PacBio HiFi reads representing ~45 fold coverage (N50 read length 14,054 bp; minimum read length 45 bp; mean read length 13,449 bp; maximum read length of 50,337 bp) based on the Genomescope2.0 genome size estimation of 1.576 Gb. Based on PacBio HiFi reads, we estimated 0.09 % sequencing error rate and 2.73% nucleotide heterozygosity rate. The Illumina sequencing yielded 203.3 million 150 bp paired end Omni-C reads.

Nuclear and mitochondrial genome assembly

We generated a *de novo* nuclear genome assembly of the California mussel (xbMytCali1) for which assembly statistics are reported in Table 2 and Figure 2B. The k-mer spectrum output shows a bimodal distribution with two major peaks, at ~21 and ~42-fold coverage, where peaks correspond to homozygous and heterozygous states respectively of a diploid species. The Omni-C contact map suggests that the primary assembly is highly contiguous (Figure 2C). The alternate assembly, which

consists of sequence from heterozygous regions, is less contiguous (Figure S1). We have deposited both the primary and alternate scaffolds to NCBI.

We generated one mitochondrial genome assembly, 16,730 bp long. The base composition of the final assembly version is A=28.41%, C=13.37%, G= 22.71%, T= 35.48%, and consists of 23 transfer RNAs and 13 protein coding genes.

Genome size estimation and quality assessment

The primary assembly consists of 176 scaffolds spanning 1.65 Gb with contig N50 of 16.32 Mb, scaffold N50 of 118 Mb, largest contig of 57.8 Mb, and largest scaffold of 142.4 Mb. The final genome size is close to the estimated values from the Genomescope2.0 k-mer spectrum. The primary assembly has a BUSCO completeness score of 86.0% using the Mollusca gene set, a per base quality (QV) of 65, a k-mer completeness of 67.2 and a frameshift indel QV of 51.43. The alternate assembly has a BUSCO completeness score of 85.20% using the Mollusca gene set, a per base quality (QV) of 59, a k-mer completeness of 68.2 and a frameshift indel QV of 51.43. The scaffold length histogram showed that the largest size differences were between scaffolds 9 and 10 (19.6 Mb) and scaffolds 12 and 13 (19.1 Mb) (Figure 2D), the latter corresponding to the split indicated by the k-means clustering, which placed scaffolds 1–12 into a cluster and scaffolds 13–20 into a second cluster.

Comparison to previously published Mytilus *genomes*

Compared to the most recent assemblies for three other *Mytilus* species (Table 3), the *M. californianus* genome assembly produced here is the most contiguous (i.e., contained in the smallest number of scaffolds), has the largest N50 value, and is the longest (1,651,966,901 bp), slightly exceeding *M. edulis* (1,651,313,236 bp). The *M. californianus* assembly also has superior BUSCO metrics for core gene completeness (86%), duplication (0.91%), fragmentation (3.3%), and missingness (10.7%). Complete single copy statistics for *M. coruscus*, *M. edulis*, and *M. galloprovincialis* are 81%, 79%, and 71% respectively; duplication statistics are 1.1%, 7.7%, and 5.4%, respectively; fragmentation statistics are 3.4%, 4%, and 4.8%, respectively; missingness statistics are 15.6%, 16.8%, and 24.7%, respectively (Table 3).

Discussion

The ecological and economic value of *Mytilus* species across the globe has motivated generation of multiple genomic resources for this genus (Murgarella et al. 2016, Yang et al. 2021, BioProject: PRJNA740305). These resources have improved in quality with advances in sequencing and assembly algorithms as can be seen comparing the first mussel genome for *M. galloprovincialis* produced with Illumina short reads (1.74 million scaffolds, N50 = 2651 bp, Murgarella et al. 2016), to the current *M. galloprovincialis* assembly produced with PacBio HiFi long reads and Illumina short reads (10,777

scaffolds, N50 = 32.14 Mb , Gerdol et al 2020). A key advance has been scaffolding assemblies with proximity data from Hi-C or Omni-C libraries, which can greatly increase contiguity. For example, for *M. coruscus*, a genome assembled from Oxford Nanopore Technology (ONT) long reads and Illumina short reads by Li et al. (2020) is contained in 10,484 scaffolds with an N50 of 898 Kb while a genome produced by Yang et al. (2021) using ONT, Illumina, and Hi-C scaffolding reduced scaffold number to 4,434 with a 99 Mb N50. Interestingly, in addition to improvements in contiguity and completeness (as determined by BUSCO metrics), assemblies scaffolded with Omni-C or Hi-C reads have less variation in assembly size (1.57-1.65Gb) than those not scaffolded with proximity data (1.28-1.9Gb), with a possible explanation being that proximity data help resolve highly repetitive areas of the genome, leading to more accurate and precise assembly sizes.

Chromosome-scale reference genomes are powerful tools because their contiguity and completeness provide more power to test important ecological and evolutionary hypotheses than a genome assembly that is fragmented and missing genes or other key genomic features. Previous studies using karyotyping have shown that *Mytilus* mussels, including *M. californianus*, *M. edulis*, *M. galloprovincialis*, and *M. trossulus* have 14 chromosomes (Ahmed and Sparks 1970; Pérez-García et al. 2014). Yang *et al.* (2021) commensurately found 90.9% of *M. coruscus* genome sequence scaffolds in their assembly mapped to 14 chromosomes based on Hi-C proximity data. The Omni-C proximity data we produced here also suggests 14 chromosomes in *M. californianus* (i.e., 14 major bins along the diagonal containing each proximity read and its mate in Figure 2C); k-means clustering suggests a similar number of chromosomes (12; Figure 2d) though the efficacy of this approach may be influenced by the assembly. Mollusc genomes are known to be highly repetitive (Murgarella et al. 2016) and heterozygous (Koehn and Gaffney 1984, Diz et al. 2008), which complicates the assembly process, and likely explains why we recovered more than 14 genomic scaffolds. Regardless, the assembly we produced here is the most complete of the *Mytilus* species available and is a powerful resource for comparative and population genomics.

While mussels have biparental inheritance of the mitochondrial genome (Ladoukakis et al. 2002, Mizi et al. 2005) and we therefore would expect to assemble two scaffolds representing the maternal and paternal contributions (Murgarella et al. 2016), we generated only one. Assembling phased genome assemblies is a bioinformatic challenge (Chin et al. 2016, Mostovoy et al. 2016), particularly when heterozygosity is high, parental sequences are not available, and an intrinsic part of pipelines is to purge duplicates. Our single mitochondrial genome assembly therefore may represent either only one of the two genomes or a chimera of the two parental contributions (Table 2). Future research efforts should engage with developing tools for better resolving maternal and paternal assemblies individually for mitochondrial, as well as nuclear, genomes.

Notwithstanding the preceding caveats, the new reference genome for *M. californianus* will facilitate multiple fields of study. For example, understanding the structure and resilience of the byssal threads may be enhanced by discovering genes encoding the formation of the threads, in a manner paralleling the genome-enabled analyses of structural genes in green mussels, *Perna viridis* (Inoue et

al 2021). Additionally, the reference genome can be coupled with comparative ecological studies to give a well-rounded understanding of how functional traits diverged in different environments, e.g. Pearce and LaBarbera (2009) showed that epifaunal species have thicker and more extensive byssal threads than infaunal species, suggesting a correlation with life habits between the two groups of organisms. Furthermore, analyzing the *Mytilus* genome can expand our knowledge of stress response and immune defense in bivalves, for example elucidating the unique bivalve gene families involved in heat shock proteins (Takeuchi et al 2016).

Reference genomes also can be useful in multiple applied contexts. First, for example, genomic resources could facilitate husbandry of *M. californianus*, which has not been substantially developed for aquaculture unlike many other *Mytilus* species. *Mytilus galloprovincialis*, the first marine mussel genome to be sequenced (Murgarella et al. 2016), has great value in its native Mediterranean Sea where it constitutes 50% of global EU aquaculture in weight (Robert et al. 2013), but incurs costs as an invasive species in many other parts of the world (Brady and Somero 2006). *Mytilus coruscus* is economically valuable and popular in Asian cuisine due to its high nutritional content (Li et al. 2020; Zhang et al. 2020). The *M. edulis* genome was sequenced for the Prince Edward Island growers to develop tools to implement a breeding program to help with the declining population (https://genomecanada.ca/project/breeding-better-blue-mussels-mytilus-edulis-developing-genomic-tools-implementation-modern-and/). Despite growing larger in length and producing twice as much meat as *M. edulis* in exploratory aquaculture studies (Yamada and Dunham 1989), there is a gap in the literature when it comes to aquaculture studies for *M. californianus* compared to its sister taxa, which this genome may help redress. Second, with biodiversity conservation as a motivator of the CCGP project, the *M. californianus* genome can act as a foundation for future work understanding the genetic diversity and population connectivity for planning marine protected areas (MPAs) and MPA networks (Jeffery et al 2022). This reference genome, coupled with future population genomics study, will deepen our understanding of the evolutionary history of this species and give us a better understanding of how the population is structured, providing a foundation for future genomic studies on ecosystem engineers across the west coast of North America.

Funding

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

Acknowledgements

PacBio Sequel II library prep and sequencing were carried out at the DNA Technologies and Expression Analysis Cores at the UC Davis Genome Center, supported by NIH Shared Instrumentation Grant 1S10OD010786-01. Deep sequencing of Omni-C libraries used the Novaseq S4 sequencing platforms at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 OD018174 Instrumentation Grant. We thank Courtney Miller for help assembling tables 1 and 2. Collections were completed under California Department of Fish and Wildlife permit #D-0023037096-0 and California Department of Parks and Recreation permit #20-820-04. Partial support was provided by Illumina for Omni-C sequencing. Part of this research was conducted using the MERCED cluster (NSF- MRI, #1429783) at the Cyberinfrastructure and Research Technologies

(CIRT) at University of California, Merced.

Data Availability

Data generated for this study are available under NCBI BioProject PRJNA777198. Raw sequencing data for sample M0D057914Y (NCBI BioSample SAMN24505264) are deposited in the NCBI Short Read Archive (SRA) under SRR18000156 for PacBio HiFi sequencing data and SRR18000154-55 for Omni-C Illumina Short read sequencing data. GenBank accessions for both primary and alternate assemblies are GCA_021869535.1 and GCA_021869935.1; and for genome sequences JAKFGE000000000 and JAKFGF000000000. The GenBank organelle genome assembly for the mitochondrial genome is CM038905.1. Assembly scripts and other data for the analyses presented can be found at the following GitHub repository: www.github.com/ccgproject/ccgp_assembly

References

Allio R, Schomaker-Bastos A, Romiguier J, Prosdocimi F, Nabholz B, & Delsuc, F. 2020. MitoFinder: Efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics. Molecular Ecology Resources 20:892–905. doi:10.1111/1755-0998.13160.

Ahmed M and Sparks AK. 1970. Chromosome number, structure and autosomal polymorphism in the marine mussels *Mytilus edulis* and *Mytilus californianus*. Biological Bulletin 138:1-13.

Brady CE, Somero GN. 2006. Ecological gradients and relative abundance of native (*Mytilus trossulus*) and invasive (*Mytilus galloprovincialis*) blue mussels in the California hybrid zone. Marine Biology 148:1249–1262. https://doi.org/10.1007/s00227-005-0177-0.

Bruno JF, Stachowicz JJ, and Bertness MD. 2003. Inclusion of facilitation into ecological theory. Trends in Ecology and Evolution 18:119–125. doi:10.1016/S0169-5347(02)00045-9.

Challis, R., Richards, E., Rajan, J., Cochrane, G., & Blaxter, M. (n.d.). BlobToolKit – Interactive quality assessment of genome assemblies. doi: 10.1101/844852

Chin C, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods 13, 1050–1054. doi:10.1038/nmeth.4035.

Dayton, P. 1972. Toward an understanding of community resilience and the potential effects of enrichment to the benthos at McMurdo Sound, Antarctica. In B. Parker [ed.], Proceedings of the colloquium on conservation problems in Antarctica. Allen Press.

DeBiasse MB, Schiebelhut LM, Escalona M, Beraut E, Fairbairn C, Marimuthu MPA, Nguyen O, Sahasrabudhe R, Dawson MN. 2020. A chromosome-level reference genome for the giant pink sea star, *Pisaster brevispinus*, a species severely impacted by wasting. Journal of Hereditary, submitted.

Diz AP and Presa P. 2008. Regional patterns of microsatellite variation in *Mytilus galloprovincialis* from the Iberian Peninsula. Marine Biology 154: 277–286. doi: 10.1007/s00227-008-0921-3.

Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, Rosani U, Venier P, Naranjo-Ortiz MA, Murgarella M, Greco S, Balseiro P, Corvelo A, Frias L, Gut M, Gabaldón T, Pallavicini A, Canchaya C, Novoa B, Alioto TS, Posada D, Figueras A. 2020. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol 21(1):275. doi: 10.1186/s13059-020-02180-3.

Gutiérrez JL, Jones CG, Strayer DL, Iribarne OO. 2003. Mollusks as ecosystem engineers: the role of shell production in aquatic habitats. Oikos 101:79–90.

Inoue K, Yoshioka Y, Tanaka H, Kinjo A, Sassa M, Ueda I, Shinzato C, Toyoda A, Itoh T. 2021. Genomics

and transcriptomics of the green mussel explain the durability of its byssus. Scientific Reports 11:5992. doi: 10.1038/s41598-021-84948-6.

Holten-Andersen N, Zhao H, Waite JH. 2009. Stiff coatings on compliant biofibers: the cuticle of *Mytilus californianus* byssal threads. Biochemistry 48(12):2752-9. doi: 10.1021/bi900018m.

Jeffery NW, Lehnert SJ, Kess T, Layton KKS, Wringe BF and Stanley RRE. 2022. Application of omics tools in designing and monitoring marine protected areas for a sustainable blue economy. Frontiers in Genetics 13:886494. doi: 10.3389/fgene.2022.886494

Jones CG, Lawton JH, and Shachak M. 1994. Organisms as ecosystem engineers. Oikos 69:373–386. doi:10.2307/ 3545850.

Koehn RK and Gaffney PM. 1984. Genetic heterozygosity and growth rate in *Mytilus edulis*. Marine Biology 82:1–7. doi: 10.1007/BF00392757.

Kinoshita T and Seki M. 2014. Epigenetic memory for stress response and adaptation in plants. Plant and Cell Physiology 55:1859-1863. doi: 10.1093/pcp/pcu125.

Ladoukakis ED, Saavedra C, Magoulas A, Zouros E. 2002. Mitochondrial DNA variation in a species with two mitochondrial genomes: the case of *Mytilus galloprovincialis* from the Atlantic, the Mediterranean and the Black Sea. Molecular Ecology 11.4:755-769. doi: 10.1046/j.1365-294x.2002.01473.x.

Li R, Zhang W, Lu J, Zhang Z, Mu C, Song W, Migaud H, Wang C, Bekaert M. 2020. The whole-genome sequencing and hybrid assembly of *Mytilus coruscus*. Frontiers in Genetics 11:1664-8021. doi:10.3389/fgene.2020.00440.

Lin M, Escalona M, Sahasrabudhe R, Nguyen O, Beraut E, Buchalski MR, Wayne RK. 2022. A reference genome assembly of the bobcat, *Lynx rufus*. Journal of Heredity, esac031. doi:10.1093/jhered/esac031.

Mizi A, Zouros E, Moschonas N, Rodakis GC. 2005. The complete maternal and paternal mitochondrial genomes of the Mediterranean mussel *Mytilus galloprovincialis*: implications for the doubly uniparental inheritance mode of mtDNA. Molecular Biology and Evolution 22.4:952-967.

Mostovoy Y, Levy-Sakin M, Lam J, Lam ET, Hastie AR, Marks P, Lee J, Chu C, Lin C, Džakul Z, Cao H, Schlebusch SA, Giorda K, Schnall-Levin M, Wall JD, Kwok PY. 2016. A hybrid approach for de novo human genome sequence assembly and phasing. Nature Methods 13:587–590. doi: 10.1038/nmeth.3865.

Murgarella M, Puiu D, Novoa B, Figueras A, Posada D, and Canchaya C. 2016. A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. PLoS ONE 11(3):e0151561. doi:10.1371/journal.pone.0151561.

Nishimura O, Hara Y, Kuraku S. 2019. Evaluating genome assemblies and gene models using gVolante.

Methods Molecular Biology 1962:247-256. doi: 10.1007/978-1-4939-9173-0_15.

Paine RT 1966. Food web complexity and species diversity. The American Naturalist 100:65–75.

Paine RT 1994. Marine rocky shores and community ecology: an experimentalist's perspective. Ecology Institute, Oldendorf, Luhe, Germany.

Pearce T and Labarbera M. 2009. A comparative study of the mechanical properties of Mytilid byssal threads. Journal of Experimental Biology 1442-8. doi: 10.1242/jeb.025544.

Pérez-García C, Morán P, & Pasantes JJ. 2014. Karyotypic diversification in *Mytilus* mussels (Bivalvia: Mytilidae) inferred from chromosomal mapping of rRNA and histone gene clusters. BMC genetics 15:84. doi:10.1186/1471-2156-15-84.

Place SP, O'Donnell MJ, and Hofmann GE. 2008. Gene expression in the intertidal mussel *Mytilus californianus*: physiological response to environmental factors on a biogeographic scale. Marine Ecology Progress Series 356:1-14.

Rhie A, Walenz BP, Koren S, & Phillippy AM. 2021. 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

Riginos C, Cunningham CW 2004. Local adaptation and species segregation in two mussel (*Mytilus edulis* x *Mytilus trossulus*) hybrid zones. Molecular Ecology 14(2):381-400. doi: 10.1111/j.1365-294X.2004.02379.x.

Robert R, Sanchez JL, Perez-Paralle L. 2013. A glimpse on the mollusc industry in Europe. Aquaculture 38(1):5-11.

Sagarin, RD and Somero GN 2006. Complex patterns of expression of heat-shock protein 70 across the southern biogeographical ranges of the intertidal mussel Mytilus californianus and snail Nucella ostrina. Journal of Biogeography, 33: 622-630. doi:10.1111/j.1365-2699.2005.01403.x.

Savolainen O, Lascoux M & Merilä J. 2013. Ecological genomics of local adaptation. Nature Review Genetics 14, 807–820. https://doi.org/10.1038/nrg3522.

Seppey M, Manni M, & Zdobnov EM. 2019. BUSCO: Assessing genome assembly and annotation completeness. Methods in Molecular Biology 227–245. doi: 10.1007/978-1-4939-9173-0_14

Shaffer HB, Toffelmier Corbett-Detig RB, Escalona M, Erickson B, Fiedler P, Gold M, Harrigan RJ, Hodges S, Luckau TK, Miller C, Oliveira DR, Shaffer KE, Shapiro B, Sork VL, and Wang IJ. 2022. Landscape genomics to enable conservation actions: The California Conservation Genomics Project. Journal of

Heredity, in press.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, & Zdobnov, EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212. doi: 10.1093/bioinformatics/btv351.

Śmietanka B, Burzyński A & Wenne R. 2010. Comparative Genomics of Marine Mussels (*Mytilus spp.*) Gender Associated mtDNA: Rapidly Evolving atp8 . Journal of Molecular Evolution 71:385–400. https://doi.org/10.1007/s00239-010-9393-4.

Soot-Ryen T. 1955. A report on the family Mytilidae (Pelecypoda). Allan Hancock Pacific Expeditions 20:1-174.

Springer SA and Crespi BJ. 2006. Adaptive Gamete-Recognition Divergence in a Hybridizing *Mytilus* Population. Evolution 61(4):772–83.

Takeuchi T, and Koyanagi R, Gyoja F, Kanda M, Hisata K, Fujie M, Goto H, Yamasaki S, Nagai K, Morino Y, Miyamoto H, Endo K, Endo H, Nagasawa H, Kinoshita S, Asakawa S, Watabe S, Satoh N, Kawashima, T. 2016. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. Zoological Letters 2:3. doi: 10.1186/s40851-016-0039-2.

Todd BD, Jenkinson TS, Escalona M, Beraut E, NguyenO, Sahasrabudhe R, Scott PA, Toffelmier E, Wang IJ, Shaffer HB. 2022. Reference genome of the northwestern pond turtle, *Actinemys marmorata*. Journal of Heredity, in press.

Tracey ML, Bellet NF & Gravem, CD. 1975. Excess allozyme homozygosity and breeding population structure in the Mussel *Mytilus californianus*. Marine Biology 32:303–311. https://doi.org/10.1007/BF00399209.

Waite JH. 2017. Mussel adhesion—Essential footwork. Journal of Experimental Biology 220:517–530. doi: 10.1242/jeb.134056.

Yamada SB and Dunham JB. 1989. *Mytilus californianus*, a new aquaculture species? Aquaculture 81(3-4): 275-284. doi:10.1016/0044-8486(89)90152-X.

Yang JL, Feng DD, Liu J, Xu JK, Chen K, Li YF, Zhu YT, Liang X, Lu Y. 2021. Chromosome-level genome assembly of the hard-shelled mussel *Mytilus coruscus*, a widely distributed species from the temperate areas of East Asia. Gigascience 10(4):giab024. doi: 10.1093/gigascience/giab024.

Zhang W, Li R, Chen X, Wang C, Gu Z, Mu C, Song W, Zhan P, Huang J. 2020. Molecular identification reveals hybrids of *Mytilus coruscus* × *Mytilus galloprovincialis* in mussel hatcheries of China. Aquaculture International 28:85–93.doi: 10.1007/s10499-019-00445-8.

Tables and Figures Captions

Table 1: Assembly Pipeline and Software Usage. Software citations are listed in the text. § Options detailed for non-default runs.

Table 2: Sequencing and assembly statistics, and accession numbers. ** Assembly quality code x.y.Q.C derived notation, from (Rhie et al. 2021). x = log10[contig NG50]; y = log10[scaffold NG50]; Q = Phred base accuracy QV (Quality value); C = % genome represented by the first 'n' scaffolds, following a known karyotype of 2n=28. Quality code for all the assembly denoted by primary assembly (xbMytCali1.0.p). BUSCO Scores. (C)omplete and (S)ingle; (C)omplete and (D)uplicated; (F)ragmented and (M)issing BUSCO genes. n, number of BUSCO genes in the set/database. Bp: base pairs*

§ Read coverage and NGx statistics have been calculated based on the estimated genome size of 1.576 Gb.

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

Table 3: BUSCO scores for *Mytilus californianus* compared with *M. coruscus, M. galloprovincialis,* and *M. edulis*. *Version of PacBio sequencing chemistry not reported

Figure 1. The distribution of the California ribbed mussel, *Mytilus californianus*. (A) World map (inset) and continental map showing the reported geographic range of *M. californianus*, from the Aleutian Islands to Isla Socorro, Mexico (Soot-Ryen 1955). Other authors have since reported a narrower distribution, such as from Baja California to British Columbia (Sagarin and Somero 2006). Yellow star indicates the geographic location of the sample collected and used to generate the genome. (B) A bed of *M. californianus* mussels at Duxbury Reef, Marin County, California. Note the other species that inhabit the ribbed mussel beds such as the acorn and stalked barnacle. Globe image in panel A taken from https://www.clipsafari.com/clips/o29983-globe-showing-north-america. Panel B photo credit: Michael N Dawson.

Figure 2. Visual overview of genome assembly metrics. (A) K-mer spectra output generated from PacBio HiFi data without adapters using GenomeScope2.0. The bimodal pattern observed corresponds to a diploid genome. K-mers covered at lower coverage but higher frequency correspond to differences between haplotypes, whereas the higher coverage but lower frequency k-mers correspond to the similarities between haplotypes. The pattern observed corresponds to a k-mer profile for a highly heterozygous species. (B) BlobToolKit Snail plot showing a graphical representation of the quality metrics presented in Table 2 for the *Mytilus californianus* primary assembly (xbMytCali1). The plot circle represents the full size of the assembly. From the inside-out, the central plot covers length-related metrics. The red line represents the size of the longest scaffold; all other scaffolds are arranged in size-order moving clockwise around the plot and drawn in gray starting from the outside of the central plot. Dark and light orange arcs show the scaffold N50 and scaffold N90 values. The central light gray spiral shows the cumulative scaffold count with a white line at each order of magnitude. White regions in this area reflect the proportion of Ns in the assembly The dark vs. light blue area around it shows mean, maximum and minimum GC vs. AT content at 0.1% intervals (Challis et al. 2020). (C) Omni-C contact maps for the primary genome assembly generated with PretextSnapshot. Omni-C contact maps translate proximity of genomic regions in 3-D space to contiguous linear organization. Each cell in the contact map corresponds to sequencing data supporting the linkage (or join) between two such regions. Scaffolds are separated by black lines, wherein higher density of black lines corresponds to higher levels of fragmentation. (D) Histogram of the 20 longest scaffold reads for *Mytilus californianus*. Scaffold size is given in megabase pairs (Mb).

Table 1

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 (HiC mode, --primary, p_ctg and a_ctg output)	0.16.1-r375
Remove low-coverage, duplicated contigs	purge_dups	1.2.6
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 20e2769
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

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

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 mollusca)	5.0.0
	Merqury	1
Contamination screening		
Local alignment tool	BLAST+	2.10
General contamination screening	BlobToolKit	2.3.3
Comparison analysis		
Assembly completeness	BUSCO (https://gvolante.riken.jp/)	5.0.0

Table 2

Bio Projects & Vouchers	CCGP NCBI BioProject			PRJNA720569
	Genera NCBI BioProject			PRJNA765636
	Species NCBI BioProject			PRJNA777198
	NCBI BioSample			SAMN24505264
	Specimen identification			M0D057914Y
	NCBI Genome accessions		Primary	Alternate
	Assembly accession		GCA_021869535.1	GCA_021869935.1
	Genome sequences		JAKFGE000000000	JAKFGF000000000
Genome Sequence	PacBio HiFi reads	Run	3 PACBIO_SMRT (Sequel II), 5.8 M spots, 81.4 G bases, 48.5 Gb	
		Accession	SRR18000156	
	Omni-C Illumina reads	Run	2 Illumina HiSeq X Ten runs: 203.3 M spots, 61.4 G bases, 20.3 Gb	
		Accession	SRR18000154-55	
Genome	Assembly identifier (Quality code *)			xbMytCali1 (7.7.Q60.C86)

Assembly Quality Metrics	HiFi Read coverage §		45.05X
		Primary	Alternate
	Number of contigs	498	38,455
	Contig N50 (bp)	16,323,199	167,345
	Contig NG50 (bp)	17,177,226	251,220
	Longest Contigs (bp)	57,759,394	3,992,896
	Number of scaffolds	176	38,315
	Scaffold N50 (bp)	117,871,512	169,002
	Scaffold NG50 (bp)	120,330,192	253,008
	Largest scaffold (bp)	142,435,203	3,992,896
	Size of final assembly (bp)	1,651,966,901	2,213,012,655
	Gaps per Gbp (#Gaps)	37 (324)	63 (140)
	Indel QV (Frame shift)	51.43139759	51.43139759
	Base pair QV	65.0891	58.9335
			Full assembly = 60.6658
	k-mer completeness	67.2429	68.2222
			Full assembly = 92.3895

BUSCO completeness		**C**	S	D	F	M
(mollusca) n=5,295	P‡	86.00%	85.10%	0.90%	3.30%	10.70%
	A‡	85.20%	80.00%	5.20%	4.50%	10.30%
Organelles		1 Complete mitochondrial sequence				CM038905.1

Table 3

	Mytilus californianus	*Mytilus coruscus*	*Mytilus edulis*	*Mytilus galloprovincialis*
Citation	This paper	Yang et al. 2021	Unpublished	Gerdol et al. 2020
GenBank ID	GCA_021869935.1	GCA_017311375.1	GCA_019925275.1	GCA_900618805.1
Sequencing method/technology	PacBio HiFi, Omni-C	Oxford Nanopore, Illumina, Hi-C	PacBio*, Omni-C	PacBio HiFi, Illumina
Assembly length (Gb)	1.65	1.57	1.65	1.28
Sequences	176	4434	1119	10,577
GC-content	32.57	32.45	32.3	32.14
N50 sequence length (Mb)	118	99.5	11.65	0.21
Complete single copies	4553 (86%)	4288 (81%)	4191 (79%)	3735 (71%)
Complete + partial single copies	4730 (89%)	4468 (84%)	4403 (83%)	3988 (75%)
BUSCO duplicated genes	48 (0.91%)	58 (1.1%)	408 (7.7%)	286 (5.4%)
BUSCO fragmented genes	175 (3.3%)	180 (3.4%)	212 (4%)	254 (4.8%)
BUSCO missing genes	567 (10.7%)	827 (15.6%)	892 (16.8%)	1307 (24.7%)

Figure 1.

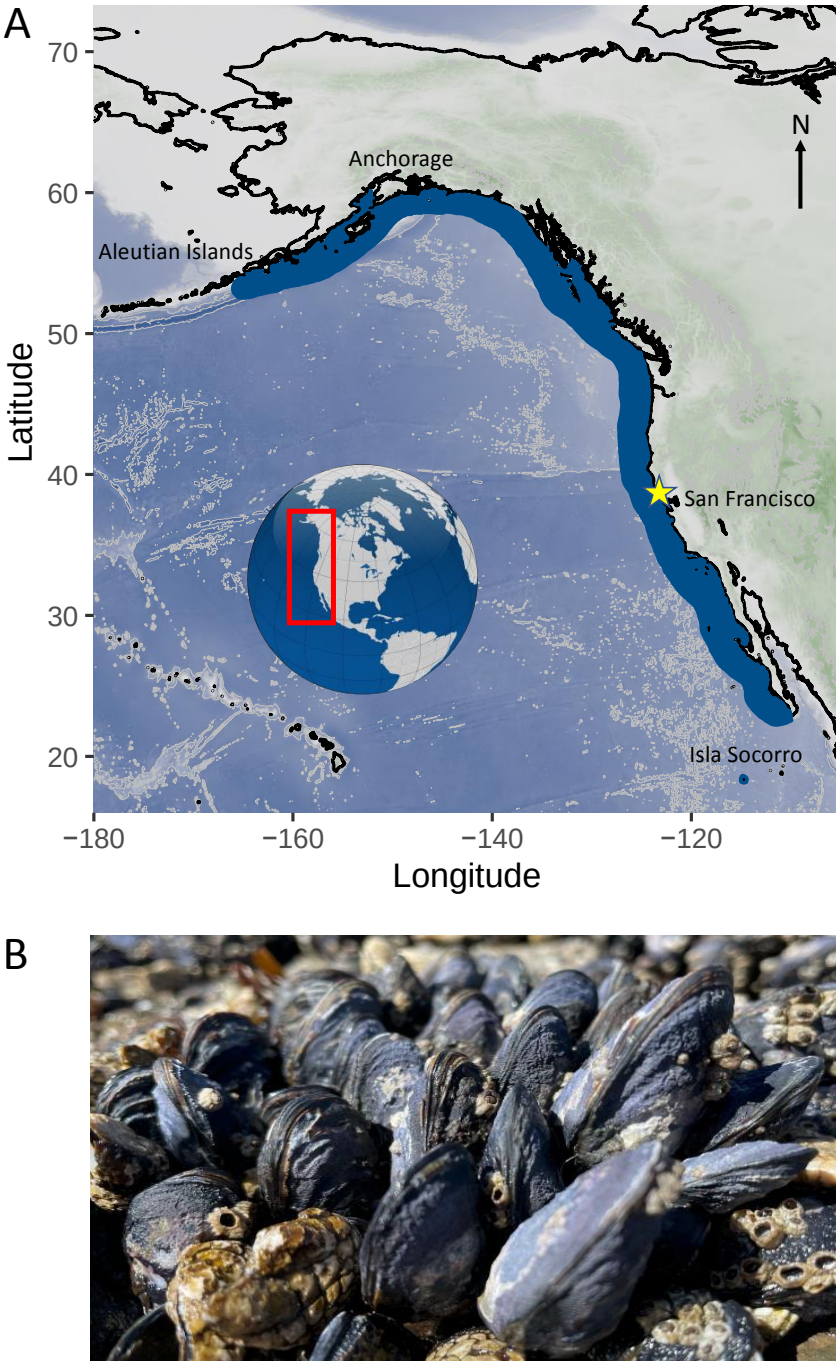

Figure 2.