UC Santa Cruz

UC Santa Cruz Previously Published Works

Title

A high-quality reference genome of the kelp surfperch, Brachyistius frenatus (Embiotocidae), a wide-ranging Eastern Pacific reef fish with no pelagic larval stage

Permalink

https://escholarship.org/uc/item/0nj8344h

Journal

Journal of Heredity, 114(4)

ISSN

0022-1503

Authors

Toy, Jason A

Bernardi, Giacomo

Publication Date

2023-06-22

DOI

10.1093/jhered/esad009

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Journal of Heredity, 2023, **114**, 404–409
https://doi.org/10.1093/jhered/esad009
Advance access publication 15 February 2023
Genome Resources

American Genetic Association

OXFORD

Genome Resources

A high-quality reference genome of the kelp surfperch, *Brachyistius frenatus* (Embiotocidae), a wide-ranging Eastern Pacific reef fish with no pelagic larval stage

Jason A. Toy, Giacomo Bernardi

Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, United States

Address correspondence to J.A. Toy at the address above, or e-mail: jasonatoy@gmail.com.

Corresponding Editor: Elizabeth Alter

Abstract

The surfperches (family Embiotocidae) are a unique group of mostly marine fishes whose phylogenetic position within the Ovalentaria clade (Percomorpha) is still unresolved. As a result of their viviparity and lack of a dispersive larval stage, surfperches are an excellent model for the study of speciation, gene flow, and local adaptation in the ocean. They are also the target of an immensely popular recreational fishery. Very few high-quality molecular resources, however, are available for this group and only for a single species. Here, we describe a highly complete reference genome for the kelp surfperch, *Brachyistius frenatus*, assembled using a combination of short-read (Illumina, ~47× coverage) and long-read (Oxford Nanopore Technologies, ~27× coverage) sequencing. The 596 Mb assembly has a completeness level of 98.1% (BUSCO), a contig N50 of 2.6 Mb (n = 56), and a contig N90 of 406.6 kb (n = 293). Comparative analysis revealed a high level of synteny between *B. frenatus* and its close relative, *Embiotoca jacksoni*. This assembly will serve as a valuable molecular resource upon which future evolutionary dynamics research will build, such as the investigation of local adaptation and the genomic potential for climate adaptation in wild populations.

Keywords: comparative genomics, de novo assembly, nanopore sequencing, Ovalentaria, Percomorpha, viviparous

Introduction

The surfperches (Embiotocidae) comprise a family of mostly marine, mostly Eastern Pacific fishes whose phylogenetic position within the Ovalentaria clade (a percomorph group that also includes the damselfishes, cichlids, mullets, and others) is still unresolved (Longo and Bernardi 2015; Ghezelayagh et al. 2022). Among marine fishes, the family of 23 species exhibits unusual life history traits. Surfperches undergo internal fertilization and viviparous gestation, and lack the dispersive pelagic larval stage exhibited by most other marine groups (Leis 1991). Surfperches are also notable within the Ovalentaria because although they comprise relatively few species, they have invaded a large diversity of nearshore habitats, including sandy surf zones, rocky reefs, kelp forests, seagrass beds, estuaries, and even coastal freshwater habitats (Tarp 1952), and can make up a large proportion of the fish biomass within these habitats in the Eastern Pacific (Laur and Ebeling 1983). This unique ecology makes the surfperches an excellent group in which to study a range of ecological and evolutionary topics including speciation, adaptive radiation, life history evolution, and local adaptation.

The kelp surfperch, *Brachyistius frenatus*, is a smaller species (max TL: 21.6 cm) within the Embiotocinae subfamily with a highly kelp-associated ecology (Love 2011). It is one of the widest ranging of the surfperches, occurring along the Pacific Coast of North America from at least Bahia Tortugas, Baja California Sur, MX to north of Sitka, Alaska, USA (personal observation; Love 2011). This makes it an ideal species in which to study local adaptation and gene flow in coastal marine environments. Here, we present a highly complete, highly contiguous de novo genome assembly for *B. frenatus* constructed with a combination of long-read nanopore and short-read shotgun sequencing data. In addition to its application to forthcoming continent-scale population genomics studies, this genome will serve as an important resource for future studies of comparative genomics and evolutionary dynamics in the Eastern Pacific.

Materials and methods

DNA sampling and sequencing

We collected an adult male *B. frenatus* from Stillwater Cove in Pebble Beach, CA, USA (36.564821, –121.943556) on 18 January 2021 (CDFW permit S-193170005-19318-001) via spear and kept it on ice during transport back to the UC Santa Cruz Coastal Science Campus (Santa Cruz, CA, USA). We sexed the specimen by dissection and sampled liver tissue for DNA extraction, after which the specimen was preserved in 95% ethanol. Genomic DNA was extracted from liver tissue using chloroform methods (adapted from Sambrook et

Received August 17, 2022; Accepted February 13, 2023

© The American Genetic Association. 2023.
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

al. 1989) and high molecular weight confirmed on a 0.7% agarose gel. Genomic DNA was then used to prepare libraries for both Nanopore long-read sequencing and Illumina short-read sequencing.

We prepared 2 libraries for 2 separate Nanopore sequencing runs using the SQK-LSK109 chemistry kit and protocol. Prior to preparation of the second library, we spun genomic DNA through a Covaris g-TUBE shearing tube to increase sequencer output. Libraries were sequenced on a Nanopore MinION device using 2 R9.4.1 flow cells. In total, we obtained 20.05 Gb of raw sequence data (6,782,325 raw reads).

Preparation of an Illumina sequencing library was done by Novogene Corporation Inc. (Sacramento, CA, USA) using fragmentation by sonication and the NEBNext Ultra DNA Library Prep Kit for Illumina. Library size distribution was evaluated on an Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA) and real-time PCR was used for quantification. Short-read data were obtained in 2 sequencing runs on an Illumina NovaSeq 6000 (2 × 150 bp reads). In total, we obtained 28.7 Gb of raw sequence data (191,338,318 raw reads).

Read processing and genome assembly

Software versions used in the assembly are listed in Table 1. Base calling of Nanopore reads was done using Oxford Nanopore Technologies' Guppy base caller software (v5.0.15) and the dna_r9.4.1_450bps_sup (super accurate) model with default quality filtering parameters, resulting in only reads with >Q7 quality scores (nearly all reads >Q10). Sequencing adapters were then trimmed using Porechop (v2.4; https://github.com/rrwick/Porechop) and trimmed reads quality checked with FastQC (v0.11.7; Andrews 2010). Finally, reads less than 500 bp were removed. This filtered and trimmed dataset contained 4,966,516 reads with a mean length of 3,268.7 bp and an N50 of 4,886 bp. Additional Nanopore sequencing stats are listed in (Supplementary Table S1). Illumina reads were trimmed with Trimmomatic (v0.39; parameters: LEADING:2 TRAILING:2 MINLEN:25, Bolger et al. 2014) and quality checked with FastQC.

The trimmed and filtered Nanopore reads were aligned into large contigs using the fuzzy-Bruijn graph-based assembler, wtdbg2 (v0.0, Ruan and Li 2020), and twice-polished with the same reads using minimap2 (v2.17-r941, Li 2018) for alignment and Racon (v1.4.13, Vaser et al. 2017) for consensus generation. The resulting assembly was then twice-polished with the trimmed Illumina short-read data using BWA (v0.7.17-r1188; Li and Durbin 2009) for alignment and Pilon (v1.23; Walker et al. 2014) for consensus generation. A blastn (v2.12.0) search of the polished assembly was then run against the nt database (NCBI) for use in contaminant detection via Blobtools2 (v3.0.0; Challis et al. 2020). In total, only 5 small contigs (less than 7,200 bp each; 27,511 bp total) were identified as contaminants (Proteobacteria) and removed from the assembly.

Following contamination removal, the mitochondrial genome was assembled and then mapped to the nuclear assembly to remove mitochondrial sequences. First, long and short reads were mapped to a reference mitogenome from the black surfperch, *Embiotoca jacksoni* (GenBank accession JAKOON010000230.1; Bernardi et al. 2022) using minimap2. The reads that successfully mapped were then imported into Geneious Prime (v2022.1.1, http://www.geneious.com/) for mitogenome assembly. In Geneious Prime, the long reads were mapped to the *E. jacksoni* reference mitogenome using the "Map to Reference" function and the default Geneious iterative mapper (Medium Sensitivity/Fast, default parameters). The generated consensus sequence was then used as a scaffold on which to map the paired short-read data for polishing, again using the Geneious mapper. This consensus sequence was then annotated using *E. jacksoni* reference annotations from GenBank accession NC_029362.1 (Longo et al. 2016) to identify potential spurious frameshift mutations and stop codons introduced by sequencing and/or assembly errors. We identified a total of 3 stop codons, one each in the coding sequences of the ND2, COI, and ND4 genes, which were all caused by incorrect calling of the length of mononucleotide repeats. After manual inspection of read data at each of these sites, an additional nucleotide was added to each repeat, which in each case resulted in the elimination of the stop codon through the re-shifting of the reading frame. A similar process was also used to change the only "N" in the sequence to an additional "C" at the end of a multi-C repeat (see Supplementary Methods). The final, corrected mitogenome assembly was then annotated using the MitoAnnotator pipeline (v3.74; Iwasaki et al. 2013).

In Geneious Prime, we ran a megablast search (NCBI, Altschul et al. 1990) with the completed mitogenome as query and the nuclear assembly as the database to identify regions of the nuclear assembly where mitochondrial sequence was incorrectly included. We filtered the list of contigs with

Table 1. Software names and versions used in the de novo genome assembly.

Assembly step	Software	Version
Nanopore base calling	Guppy	5.0.15
Nanopore adaptor trimming	Porechop	2.4
Illumina read trimming	Trimmomatic	0.39
De novo long-read contig assembly	Wtdgb2	0.0
Long-read contig polishing mapping	Minimap2	2.17-r941
Long-read contig polishing consensus generation	Racon	1.4.13
Short-read contig polishing mapping	BWA	0.7.17-r1188
Short-read contig polishing consensus generation	Pilon	1.23
Contaminant detection	Blobtools2	3.0.0
Completeness evaluation	BUSCO	5.2.2
QV score and k-mer completeness	Merqury	1.3
De novo repeat element identification	RepeatModeler	2.0.3
Repeat content analysis	RepeatMasker	4.1.2-p1
Filtering of mitochondrial reads	Minimap2	2.17-r941
Alignment of mitochondrial reads to reference	Geneious Prime	2022.1.1
Mitochondrial sequence consensus generation	Geneious Prime	2022.1.1
Mitochondrial genome annotation	MitoAnnotator	3.74

BLAST hits by considering for removal only those contigs where BLAST hits to mitochondrial sequence made up >20% of the contig length. All other hits were assumed to be potential NUMTs (nuclear DNA segments of mitochondrial origin). We removed mitochondrial sequence from a total of 4 small contigs (largest: 11,036 bp). In each case, the validity of the remaining contig fragments was assessed by mapping long-read data to each fragment. Fragments or portions of fragments with clear support from long-read data were kept as new contigs. In total, this process led to the complete removal of 2 contigs from the nuclear assembly, the splitting of 1 contig into 2, and the trimming of another. For details, see Supplementary Methods.

We identified repeat elements de novo by running RepeatModeler (v2.0.3; Flynn et al. 2020) on the assembled genome sequence using the -LTRStruct option. Using RepeatMasker (v4.1.2-p1; Smit et al. 2013), we then determined the repeat content of the genome by running a slow search (-s parameter) using the custom repeat library created with RepeatModeler (-lib parameter). We also used Merqury (v1.3; Rhie et al. 2020) to estimate the QV score and k-mer completeness of the assembly using a k-mer size of 21.

Synteny analysis

To compare the genome structure of *B. frenatus* to that of a close relative, an initial version of the assembly was aligned to an existing high-quality assembly for the genome of a close relative, the black surfperch, *E. jacksoni* (GCA_022577435.1; Bernardi et al. 2022), using minimap2 (Li 2018) via the D-GENIES application (Cabanettes and Klopp 2018). This mapping led to the identification of an

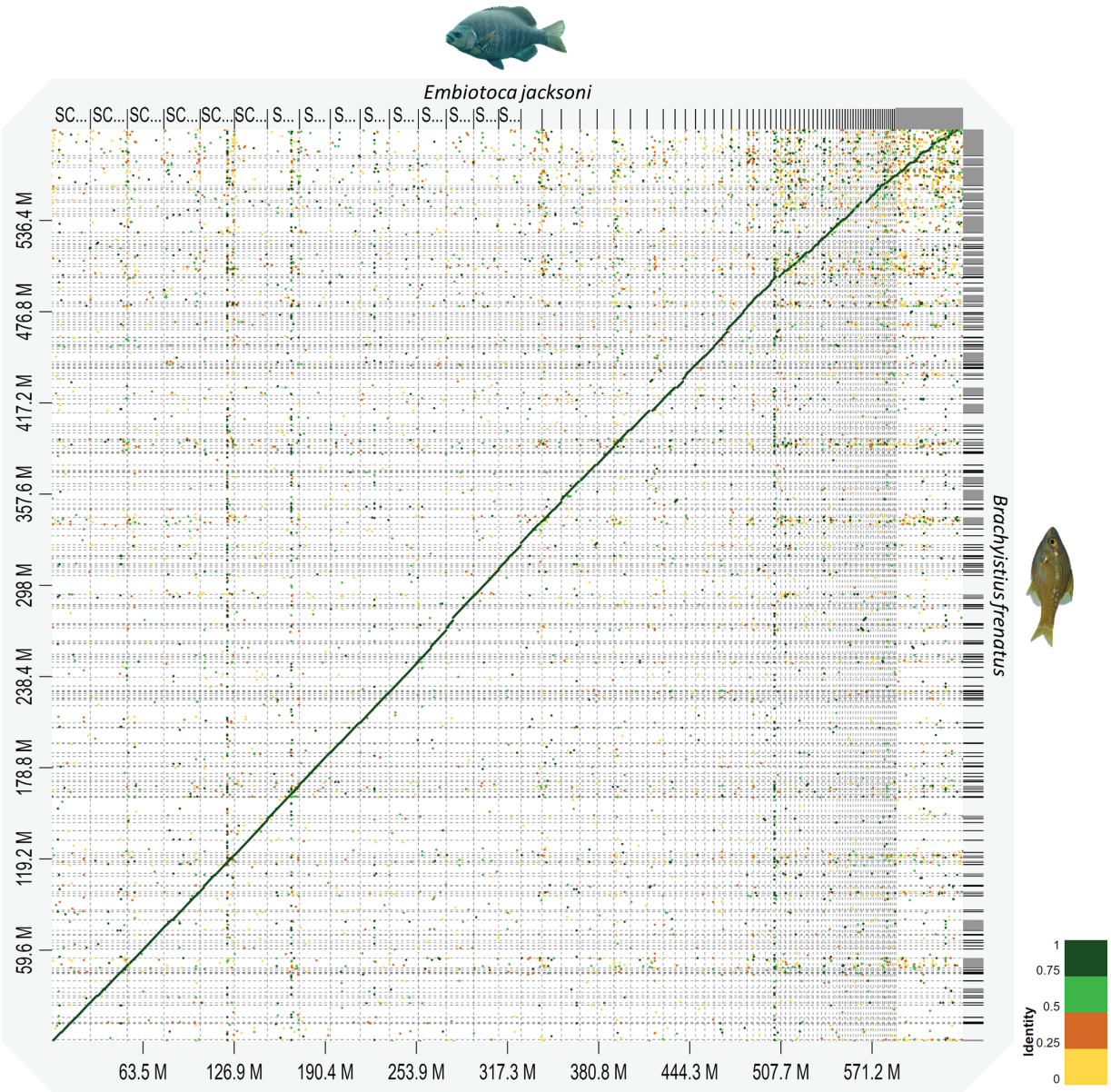

Fig. 1. Dot plot produced using D-GENIES of the alignment of the *B. frenatus* genome (vertical axis) to a reference *E. jacksoni* genome (horizontal axis; accession GCA_022577435.1). Dotted gridlines represent scaffold/contig boundaries. Darker/greener colors indicate greater sequence similarity between the query and reference sequences. *E. jacksoni* scaffolds are ordered numerically in ascending order from left to right.

apparent "translocation," which upon further investigation, had little read support and was therefore determined to be a misassembly of 2 separate contigs. This contig was therefore split at a clear break in read support (see Supplementary Methods for details). The new assembly was again mapped to the *E. jacksoni* reference using D-GENIES (Fig. 1).

An alternate version of the assembly was also created by further scaffolding the draft *B. frenatus* assembly with RagTag (v2.1.0; Alonge et al. 2022), using the *E. jacksoni* assembly as a reference. This scaffolding reduces the number of contigs from 1,004 to 355. This scaffolding resulted in the successful alignment of 594,362,913 bp of the *B. frenatus* assembly to the *E. jacksoni* reference (99.7% alignment rate).

Results

A highly complete *B. frenatus* reference genome assembly

After contamination screening and mitogenome assembly, the final genome assembly has a size of 595,951,806 bp distributed across 1,004 contigs (including the mitogenome). This indicates an average long-read coverage of ~27× and an average short-read coverage of ~47×. The largest contig is 12.3 Mb in length, and 50% of the genome is contained in 56 contigs that are ~2.6 Mb or greater in length (contig N50, Table 2). Ninety percent of the genome is contained in 293 contigs of length 406.6 kb or greater (contig N90). The inner circle of Fig. 2 shows the size distribution of all contigs in the genome. BUSCO analysis of the genome revealed a high level of completeness, with 98.1% of orthologs identified as present and complete. Evaluation of the assembly with Merqury resulted in a k-mer completeness of 93.83% and a QV score of 38.75 (corresponding to an error rate of 0.000133466). See Table 2 for additional assembly statistics and Supplementary Fig. S1 for the copy number spectrum plot from the Merqury k-mer analysis.

The final mitochondrial assembly is 16,545 bp in length, 30 bp longer than the *E. jacksoni* reference used for assembly. Pairwise alignment of the *B. frenatus* mitogenome and reference sequence revealed a 90.2% pairwise identity between the 2 species. The base composition of the *B. frenatus* mitogenome assembly is A = 28.0%, C = 27.6%, G = 16.3%, and T = 28.1%. Annotation by MitoAnnotator identified 22 tRNAs, the 12S and 16S rRNAs, and 13 protein-coding genes.

In total, RepeatMasker identified 125,790,847 bp of repeat sequence representing 21.11% of the genome. Retroelements accounted for 3.54% of the genome and DNA transposons 4.75%. Simple repeats made up 4.29% of the genome, while low complexity regions and small RNA (rRNA and tRNA) accounted for 0.38% and 0.14%, respectively (Supplementary Table S2).

Table 2. Assembly statistics and BUSCO completeness assessment for the *B. frenatus* genome.

Assembly statistics	
Assembly	Nuclear + mitochondrial genome
Size (bp)	595,951,806
n scaffolds	1,004
Average scaffold length	593,577.50
Largest scaffold	12,326,090
N50 (bp, *n*)	2,589,815 (56)
N60	1,896,895 (83)
N70	1,185,759 (123)
N80	758,445 (185)
N90	406,612 (293)
N100	597 (1,004)
N count	0
Gaps	0
BUSCO results	
Complete	98.1% (3,572)
Complete and single copy	97.4% (3,547)
Complete and duplicated	0.7% (25)
Fragmented	0.5% (20)
Missing	1.4% (48)
Total BUSCOS searched	3,640

All "N" statistics refer to contigs and are presented in the format, "length (number of contigs of that length or greater)."

Discussion

In this study, we present the second complete reference genome for the family Embiotocidae, providing new opportunities for comparative genomic and phylogenetic analyses within and outside of this group of uncertain placement. The assembly is of high quality, contiguity (contig N50 = 2.6 Mb), and completeness (98.1%) as a result of our combination of high accuracy shotgun and long-read nanopore sequencing. With 90% of the genome contained within only 293 contiguous sequences of length 406.6 kb or greater (N90), this assembly will serve as a reliable resource for comparative genomic studies and the estimation of population genomic parameters using whole-genome resequencing data.

Overall, we found that the genome of *B. frenatus* is very similar to that of its close relative, *E. jacksoni*. The dot plot created with D-GENIES reveals a high level of synteny and sequence similarity between the 2 species (Fig. 1), indicating speciation between the 2 occurred without a major structural rearrangement. The genome assemblies are also similar in size (596 vs. 635 Mb) and GC content (41.9% vs 41.6%), as would be expected for closely related species (Bernardi et al. 2022). Moving forward, this molecular resource will serve as the critical foundation for resequencing studies focusing on the characterization of genomic diversity and gene flow, as well as the evaluation of past and potential future climate adaptation, in reef fishes along the Pacific Coast. This work will ultimately inform management and conservation efforts, such as the classification of fisheries stocks and the evaluation of MPA efficacy.

Supplementary material

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

Supplementary Table S1. Summary table of MinION sequencing reads after quality filtering, adapter trimming, and length filtering (>500 bp). Produced using NanoStat (v1.5.0).

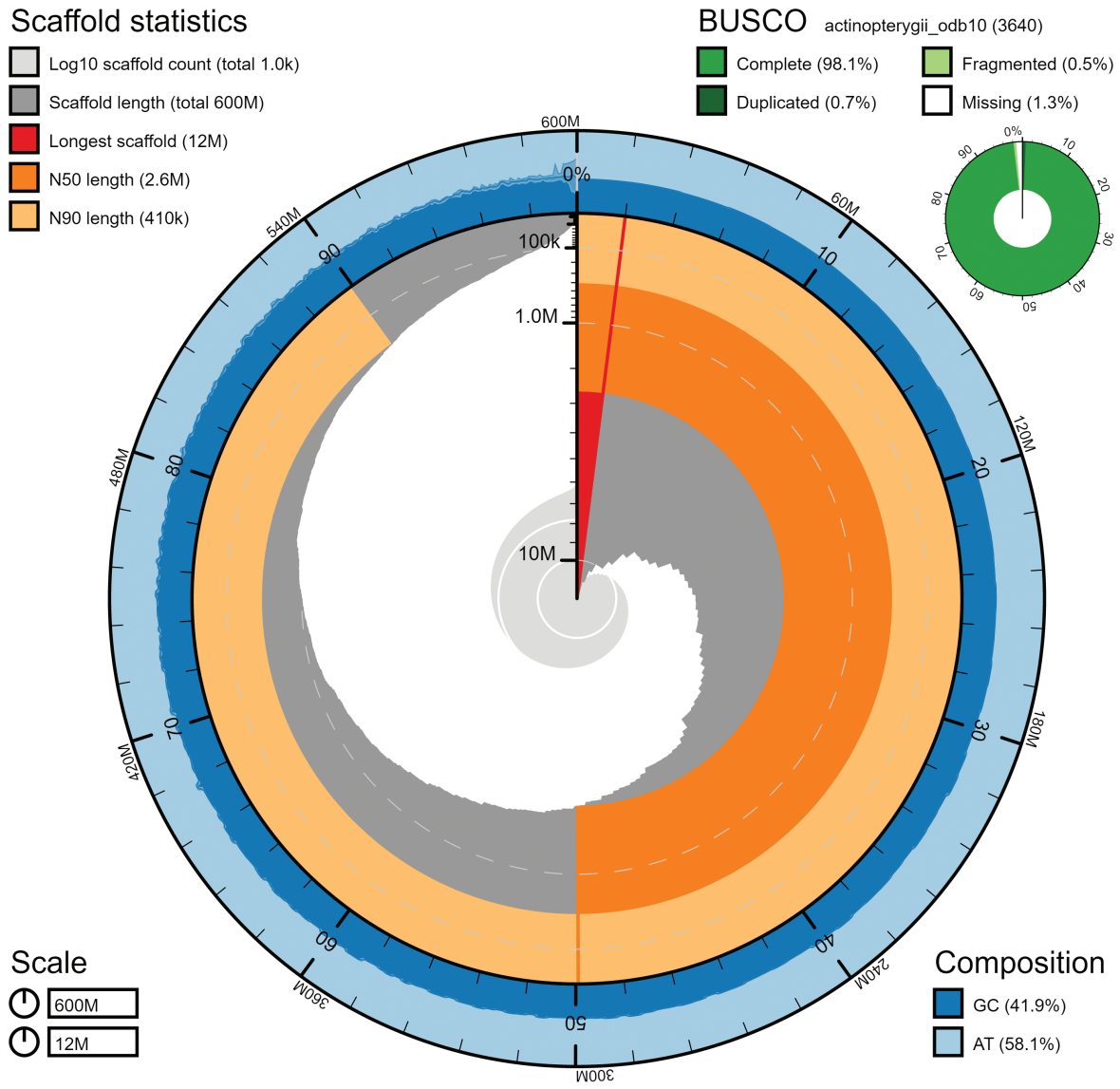

Fig. 2. Snail summary plot of the complete *B. frenatus* genome assembly produced using Blobtools2. The inner radial axis (gray) shows the length of each contig in descending order, with dark orange and light orange portions representing the N50 and N90 lengths, respectively. The dark/light blue ring shows the GC content along the length of the genome. The notched outer ring denotes the position (bp/proportion) along the genome. The inset at the top right shows the assembly completeness as assessed by BUSCO.

Supplementary Table S2. Summary of identified repeat elements. Elements were identified using RepeatModeler and repeat content summarized with RepeatMasker.

Supplementary Fig. S1. Copy number spectrum plot produced with Merqury.

Funding

Financial support for this project was provided by the Steven Berkeley Marine Conservation Fellowship (American Fisheries Society) awarded to JAT.

Acknowledgments

We would like to thank Kristy Kroeker for support during the planning and execution of the project, and Remy Gatins and Merly Escalona for valuable assistance and insight during the assembly of the genome. We also thank Rion Parsons for technical assistance, Celine de Jong for support during field collection and dissection of the sample, and Dan Wright for assistance in the laboratory.

Conflict of interest

The authors declare no competing interests.

Data availability

Raw read data for NCBI BioProject PRJNA862534, BioSample SAMN29982938 are deposited in the NCBI Short Read Archive (SRA) under SRR21095639 and SRR21095640 for Nanopore data and SRR20680795 for short read Illumina data. The GenBank accession number for the full assembly (including the mitochondrial assembly) is JANHZZ000000000 and the annotated mitochondrial genome is accessioned as OP238470. Assembly scripts can be found at https://github.com/jtoy7.

References

Alonge M, Lebeigle L, Kirsche M, Jenike K, Ou S, Aganezov S, Wang X, Lippman ZB, Schatz MC, Soyk S. Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol. 2022;23(1):258. doi:10.1186/s13059-022-02823-7

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403–410.

Andrews S. FastQC: a quality control tool for high throughput sequence data [Online]. 2010. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Bernardi G, Toy JA, Escalona M, Marimuthu MPA, Sahasrabudhe R, Nguyen O, Sacco S, Beraut E, Toffelmier E, Miller C, et al. Reference genome of the Black Surfperch, *Embiotoca jacksoni* (Embiotocidae, Perciformes), a California kelp forest fish that lacks a pelagic larval stage. *J Hered*. 2022;113(6):657–664. doi:10.1093/jhered/esac034

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114–2120.

Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*. 2018;6:e4958.

Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit—interactive quality assessment of genome assemblies. *G3*. 2020;10(4):1361–1374.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA*. 2020;117(17):9451–9457.

Ghezelayagh A, Harrington RC, Burress ED, Campbell MA, Buckner JC, Chakrabarty P, Glass JR, McCraney WT, Unmack PJ, Thacker CE, et al. Prolonged morphological expansion of spiny-rayed fishes following the end-Cretaceous. *Nat Ecol Evol*. 2022;6:1211–1220. doi:10.1038/s41559-022-01801-3

Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, Sado T, Mabuchi K, Takeshima H, Miya M, et al. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Mol Biol Evol*. 2013;30(11):2531–2540.

Laur DR, Ebeling AW. Predator-prey relationships in surfperches. In: Noakes DLG, Lindquist DG, Helfman GS, Ward JA, editors. Predators and Prey in Fishes: Proceedings of the 3rd Biennial Conference on the Ethology and Behavioral Ecology of Fishes, held at Normal, Illinois, U.S.A., May 19–22, 1981. Springer Netherlands; 1983. p. 55–67.

Leis JM. The pelagic stage of reef fishes. In: Sale PF, editor. *The ecology of fishes on coral reefs*. San Diego: Academic Press; 1991. p. 182–229.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34(18):3094–3100.

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. 2009;25(14):1754–1760.

Longo G, Bernardi G. The evolutionary history of the embiotocid surfperch radiation based on genome-wide RAD sequence data. *Mol Phylogenet Evol*. 2015;88:55–63.

Longo GC, O'Connell B, Green RE, Bernardi G. The complete mitochondrial genome of the black surfperch, *Embiotoca jacksoni*: selection and substitution rates among surfperches (Embiotocidae). *Mar Genomics*. 2016;28:107–112. doi:10.1016/j.margen.2016.03.006

Love M. *Certainly more than you want to know about the fishes of the Pacific Coast: a postmodern experience*. Really Big Press; 2011.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. 2020;21(1):245.

Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods*. 2020;17(2):155–158.

Sambrook J, Fritsch EF, Maniatis T. *Molecular cloning: a laboratory manual*. 2nd ed. New York: Cold Spring Harbor Laboratory Press; 1989.

Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013. http://www.repeatmasker.org

Tarp FH. 1952. Fish Bulletin No. 88. A revision of the family Embiotocidae (the surfperches). https://escholarship.org/uc/item/3qx7s3cn

Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res*. 2017;27(5):737–746.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 2014;9(11):e112963.