Check for updates

DATA NOTE

The genome sequence of the Greater sandeel, *Hyperoplus immaculatus* (Corbin, 1950) (Uranoscopiformes: Ammodytidae)

[version 1; peer review: 1 approved, 2 approved with reservations]

Kesella Scott-Somme[1], Rachel Brittain[1],
Marine Biological Association Genome Acquisition Lab,
Darwin Tree of Life Barcoding Collective,
Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team,
Wellcome Sanger Institute Scientific Operations: Sequencing Operations,
Wellcome Sanger Institute Tree of Life Core Informatics team,
Tree of Life Core Informatics collective, Darwin Tree of Life Consortium

[1]The Marine Biological Association, Plymouth, England, UK

v1 **First published:** 05 Nov 2025, **10**:622
https://doi.org/10.12688/wellcomeopenres.25072.1
Latest published: 05 Nov 2025, **10**:622
https://doi.org/10.12688/wellcomeopenres.25072.1

Abstract

We present a genome assembly from an individual *Hyperoplus immaculatus* (Greater sandeel; Chordata; Actinopteri; Uranoscopiformes; Ammodytidae). The genome sequence has a total length of 797.79 megabases. Most of the assembly (91.28%) is scaffolded into 24 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 16.53 kilobases. Gene annotation of this assembly on Ensembl identified 22 130 protein-coding genes. This assembly was generated as part of the Darwin Tree of Life project, which produces reference genomes for eukaryotic species found in Britain and Ireland.

Keywords

Hyperoplus immaculatus; Greater sandeel; Corbin's Ssndeel; genome sequence; chromosomal; Uranoscopiformes

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ? ✓ ?

	1	2	3
version 1 05 Nov 2025	? view	✓ view	? view

1. **Nick Jeffery**, Bedford Institute of Oceanography, Fisheries and Oceans Canada, Canada, Canada
2. **Xizhao Zhang**, Freshwater Fisheries and Germplasm Resources Utilization, Wuxi, China
3. **David Barros-García**, Centro Interdisciplinar de Investigação Marinha E Ambiental (CIIMAR/CIMAR), Matosinhos, Portugal

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Darwin Tree of Life Consortium (mark.blaxter@sanger.ac.uk)

Author roles: Scott-Somme K: Investigation, Resources; **Brittain R**: Investigation, Resources;

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by Wellcome through core funding to the Wellcome Sanger Institute (220540) and the Darwin Tree of Life Discretionary Award [218328, https://doi.org/10.35802/218328].
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2025 Scott-Somme K *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Scott-Somme K, Brittain R, Marine Biological Association Genome Acquisition Lab *et al.* **The genome sequence of the Greater sandeel, *Hyperoplus immaculatus* (Corbin, 1950) (Uranoscopiformes: Ammodytidae) [version 1; peer review: 1 approved, 2 approved with reservations]** Wellcome Open Research 2025, **10**:622 https://doi.org/10.12688/wellcomeopenres.25072.1

First published: 05 Nov 2025, **10**:622 https://doi.org/10.12688/wellcomeopenres.25072.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Actinopterygii; Actinopteri; Neopterygii; Teleostei; Osteoglossocephalai; Clupeocephala; Euteleosteomorpha; Neoteleostei; Eurypterygia; Ctenosquamata; Acanthomorphata; Euacanthomorphacea; Percomorphaceae; Eupercaria; Uranoscopiformes; Ammodytidae; *Hyperoplus*; *Hyperoplus immaculatus* (Corbin, 1950) (NCBI:txid1484357)

Background

A sandeel is a small, elongate marine fish in the family Ammodytidae. The greater sandeel *Hyperoplus immaculatus* is a sandeel of the north-east Atlantic ocean. It lives over clean sand on the continental shelf, with juveniles most often recorded close inshore. The body is long with a pointed head and it lacks the dark snout spot seen in related species. Adults reach about 35 cm standard length. Small individuals feed on zooplankton, while larger ones take small fishes, including clupeids and other sandeels. Spawning is in winter (Froese & Pauly, 2025).

The species was first described as *Ammodytes immaculatus* by Corbin in 1950 (Corbin, 1950). Its core range includes the coasts of the British Isles, the North Sea, the English Channel and the northern Bay of Biscay. Records from Portuguese continental waters in 2017–2018 extend the known southern limit and suggest recent establishment (Gomes *et al.*, 2018).

We present a chromosome-level genome sequence for *Hyperoplus immaculatus*, assembled via the Tree of Life pipeline from a specimen collected in Whitsand Bay, Cornwall, UK (Figure 1). Only five genomes are available for the family Ammodytidae. The present assembly is one of two genomes for the genus *Hyperoplus* as of October 2025 (data obtained via NCBI datasets, O'Leary *et al.*, 2024). It was generated as part of the Darwin Tree of Life Project, which aims to generate high-quality reference genomes for all named eukaryotic species in Britain and Ireland to support research, conservation, and the sustainable use of biodiversity (Blaxter *et al.*, 2022).

Figure 1. Photograph of the *Hyperoplus immaculatus* (fHypImm3) specimen used for genome sequencing.

Methods

Sample acquisition and DNA barcoding

The specimen used for genome sequencing was an adult *Hyperoplus immaculatus* (specimen ID MBA-211013-001B, ToLID fHypImm3; Figure 1), collected from Whitsand Bay, Cornwall, UK (latitude 50.3253, longitude –4.2363) on 2021-10-13. The specimen was collected by Kesella Scott-Somme and Rachel Brittain, and formally identified by Kesella Scott-Somme. The same specimen was used for RNA sequencing. Sample metadata were collected in line with the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

The initial identification was verified by an additional DNA barcoding process according to the framework developed by Twyford *et al.* (2024). A small sample was dissected from the specimen and stored in ethanol, while the remaining parts were shipped on dry ice to the Wellcome Sanger Institute (WSI) (see the protocol). The tissue was lysed, the COI marker region was amplified by PCR, and amplicons were sequenced and compared to the BOLD database, confirming the species identification (Crowley *et al.*, 2023). Following whole genome sequence generation, the relevant DNA barcode region was also used alongside the initial barcoding data for sample tracking at the WSI (Twyford *et al.*, 2024). The standard operating procedures for Darwin Tree of Life barcoding are available on protocols.io.

Nucleic acid extraction

Protocols for high molecular weight (HMW) DNA extraction developed at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory are available on protocols.io (Howard *et al.*, 2025). The fHypImm3 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the head was homogenised by cryogenic disruption using the Covaris cryoPREP® Automated Dry Pulverizer.

HMW DNA was extracted using the Automated MagAttract v2 protocol. DNA was sheared into an average fragment size of 12–20 kb following the Megaruptor®3 for LI PacBio protocol. Sheared DNA was purified by automated SPRI (solid-phase reversible immobilisation). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. Fragment size distribution was evaluated by running the sample on the FemtoPulse system. For this sample, the final post-shearing DNA had a Qubit concentration of 10.1 ng/μL and a yield of 4 040.00 ng.

RNA was extracted from head tissue of fHypImm3 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ *mir*Vana protocol. The RNA concentration was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit RNA Broad-Range Assay kit. Analysis of the integrity of the RNA was done using

the Agilent RNA 6000 Pico Kit and Eukaryotic Total RNA assay.

PacBio HiFi library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA), following the manufacturer's instructions. The kit includes reagents for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead clean-up, and nuclease treatment. Size selection and clean-up were performed using diluted AMPure PB beads (Pacific Biosciences). DNA concentration was quantified using a Qubit Fluorometer v4.0 (ThermoFisher Scientific) and the Qubit 1X dsDNA HS assay kit. Final library fragment size was assessed with the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) using the gDNA 55 kb BAC analysis kit.

The sample was sequenced using the Sequel IIe system (Pacific Biosciences, California, USA). The concentration of the library loaded onto the Sequel IIe was in the range 40–135 pM. The SMRT link software, a PacBio web-based end-to-end workflow manager, was used to set-up and monitor the run, and to perform primary and secondary analysis of the data upon completion.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen tissue from the head of the fHypImm3 sample using the Arima-HiC v2 kit (Arima Genomics). Following the manufacturer's instructions, tissue was fixed and DNA crosslinked using TC buffer to a final formaldehyde concentration of 2%. The tissue was homogenised using the Diagnocine Power Masher-II. Crosslinked DNA was digested with a restriction enzyme master mix, biotinylated, and ligated. Clean-up was performed with SPRISelect beads before library preparation. DNA concentration was measured with the Qubit Fluorometer (Thermo Fisher Scientific) and Qubit HS Assay Kit. The biotinylation percentage was estimated using the Arima-HiC v2 QC beads.

Hi-C library preparation and sequencing

Biotinylated DNA constructs were fragmented using a Covaris E220 sonicator and size selected to 400–600 bp using SPRISelect beads. DNA was enriched with Arima-HiC v2 kit Enrichment beads. End repair, A-tailing, and adapter ligation were carried out with the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs), following a modified protocol where library preparation occurs while DNA remains bound to the Enrichment beads. Library amplification was performed using KAPA HiFi HotStart mix and a custom Unique Dual Index (UDI) barcode set (Integrated DNA Technologies). Depending on sample concentration and biotinylation percentage determined at the crosslinking stage, libraries were amplified with 10–16 PCR cycles. Post-PCR clean-up was performed with SPRISelect beads. Libraries were quantified using the AccuClear Ultra High Sensitivity dsDNA Standards Assay Kit (Biotium) and a FLUOstar Omega plate reader (BMG Labtech).

Prior to sequencing, libraries were normalised to 10 ng/μL. Normalised libraries were quantified again and equimolar and/or weighted 2.8 nM pools were created. Pool concentrations were checked using the Agilent 4200 TapeStation (Agilent) with High Sensitivity D500 reagents before sequencing. Sequencing was performed using paired-end 150 bp reads on the Illumina NovaSeq 6000.

RNA library preparation and sequencing

Libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina (New England Biolabs), following the manufacturer's instructions. Poly(A) mRNA in the total RNA solution was isolated using oligo(dT) beads, converted to cDNA, and uniquely indexed; 14 PCR cycles were performed. Libraries were size-selected to produce fragments between 100–300 bp. Libraries were quantified, normalised, pooled to a final concentration of 2.8 nM, and diluted to 150 pM for loading. Sequencing was carried out on the Illumina NovaSeq 6000 to generate 150-bp paired-end reads.

Genome assembly

Prior to assembly of the PacBio HiFi reads, a database of *k*-mer counts (k = 31) was generated from the filtered reads using FastK. GenomeScope2 (Ranallo-Benavidez *et al.*, 2020) was used to analyse the *k*-mer frequency distributions, providing estimates of genome size, heterozygosity, and repeat content.

The HiFi reads were assembled using Hifiasm (Cheng *et al.*, 2021) with the --primary option. Haplotypic duplications were identified and removed using purge_dups (Guan *et al.*, 2020). The Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019), and the contigs were scaffolded in YaHS (Zhou *et al.*, 2023) with the --break option for handling potential misassemblies. The scaffolded assemblies were evaluated using Gfastats (Formenti *et al.*, 2022), BUSCO (Manni *et al.*, 2021) and MERQURY.FK (Rhie *et al.*, 2020).

The mitochondrial genome was assembled using MitoHiFi (Uliano-Silva *et al.*, 2023), which runs MitoFinder (Allio *et al.*, 2020) and uses these annotations to select the final mitochondrial contig and to ensure the general quality of the sequence.

Assembly curation

The assembly was decontaminated using the Assembly Screen for Cobionts and Contaminants (ASCC) pipeline. TreeVal was used to generate the flat files and maps for use in curation. Manual curation was conducted primarily in PretextView and HiGlass (Kerpedjiev *et al.*, 2018). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Manual corrections included one break and 6 joins. The curation process is documented at https://gitlab.com/wtsi-grit/rapid-curation. PretextSnapshot was used to generate a Hi-C contact map of the final assembly.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020) was run in a Singularity container (Kurtzer *et al.*, 2017) to evaluate *k*-mer

completeness and assembly quality for the primary and alternate haplotypes using the *k*-mer databases (k = 31) computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

The genome was analysed using the BlobToolKit pipeline, a Nextflow implementation of the earlier Snakemake version (Challis *et al.*, 2020). The pipeline aligns PacBio reads using minimap2 (Li, 2018) and SAMtools (Danecek *et al.*, 2021) to generate coverage tracks. It runs BUSCO (Manni *et al.*, 2021) using lineages identified from the NCBI Taxonomy (Schoch *et al.*, 2020). For the three domain-level lineages, BUSCO genes are aligned to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) using DIAMOND blastp (Buchfink *et al.*, 2021). The genome is divided into chunks based on the density of BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database with DIAMOND blastx. Sequences without hits are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The BlobToolKit suite consolidates all outputs into a blobdir for visualisation. The BlobToolKit pipeline was developed using nf-core tooling (Ewels *et al.*, 2020) and MultiQC (Ewels *et al.*, 2016), with containerisation through Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017).

Genome sequence report

Sequence data

PacBio sequencing of the *Hyperoplus immaculatus* specimen generated 65.85 Gb (gigabases) from 5.87 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 753.95 Mb, with a heterozygosity of 0.39% and repeat content of 29.83% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 31× coverage. Hi-C sequencing produced 150.02 Gb from 993.51 million reads, which were used to scaffold the assembly. RNA sequencing data were also generated and are available in public sequence repositories. Table 1 summarises the specimen and sequencing details.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The final assembly has a total length of 797.79 Mb in 1 567 scaffolds, with gaps, and a scaffold N50 of 31.41 Mb (Table 2).

Most of the assembly sequence (91.28%) was assigned to 24 chromosomal-level scaffolds. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3).

The mitochondrial genome was also assembled. This sequence is included as a contig in the multifasta file of the genome submission and as a standalone record.

The combined primary and alternate assemblies achieve an estimated QV of 53.6. The *k*-mer completeness is 93.04% for the primary assembly, 90.14% for the alternate haplotype, and 98.45% for the combined assemblies (Figure 4).

BUSCO v5.3.2 analysis using the actinopterygii_odb10 reference set (n = 3,640) identified 97.6% of the expected gene set

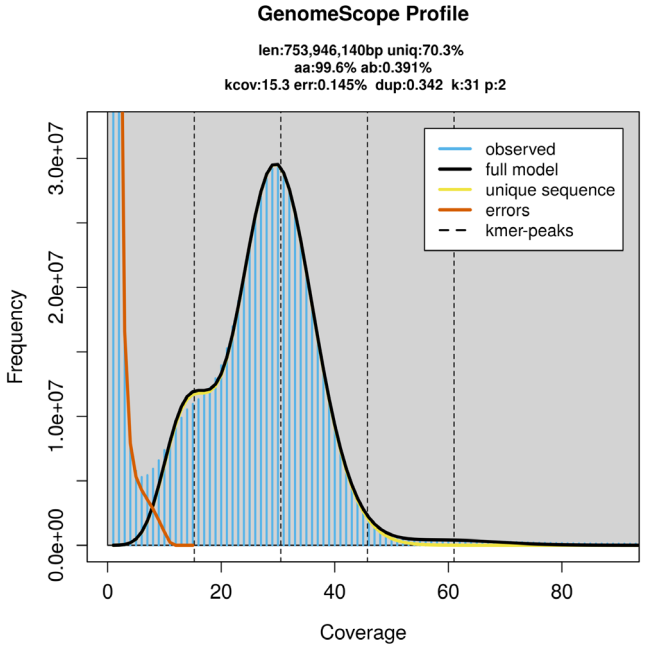

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. **Specimen and sequencing data for BioProject PRJEB58244.**

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	fHypImm3	fHypImm3	fHypImm3
Specimen ID	MBA-211013-001B	MBA-211013-001B	MBA-211013-001B
BioSample (source individual)	SAMEA13854416	SAMEA13854416	SAMEA13854416
BioSample (tissue)	SAMEA13854558	SAMEA13854558	SAMEA13854558
Tissue	head	head	head
Instrument	Sequel IIe	Illumina NovaSeq 6000	Illumina NovaSeq 6000
Run accessions	ERR10841313	ERR10851506	ERR11641116
Read count total	5.87 million	993.51 million	60.55 million
Base count total	65.85 Gb	150.02 Gb	9.14 Gb

Table 2. **Genome assembly statistics.**

Assembly name	fHypImm3.1
Assembly accession	GCA_949357725.1
Alternate haplotype accession	GCA_949358355.1
Assembly level	chromosome
Span (Mb)	797.79
Number of chromosomes	24
Number of contigs	2 755
Contig N50	1.49 Mb
Number of scaffolds	1 567
Scaffold N50	31.41 Mb
Organelles	Mitochondrion: 16.53 kb

(single = 96.8%, duplicated = 0.8%). The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for the primary assembly. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage

Table 4 lists the assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The EBP metric, calculated for the primary assembly, is **6.C.Q53**, meeting the recommended reference standard.

Genome annotation report

The *Hyperoplus immaculatus* genome assembly (GCA_949357725.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI), using the Ensembl Genebuild method. This annotation includes 68 441 transcribed mRNAs from 22 130 protein-coding and 4 575 non-coding genes. The average transcript length is 24 529.51 bp, with an average of 2.56 coding transcripts per gene and 14.11 exons per transcript. For further information about the annotation, please refer to the annotation page on Ensembl.

Wellcome Sanger Institute – Legal and Governance

The materials that have contributed to this genome note have been supplied by a Darwin Tree of Life Partner. The submission of materials by a Darwin Tree of Life Partner is subject to the **'Darwin Tree of Life Project Sampling Code of Practice'**, which can be found in full on the Darwin Tree of Life website. By agreeing with and signing up to the Sampling Code of Practice, the Darwin Tree of Life Partner agrees they will meet the legal and ethical requirements and standards set out within this document in respect of all samples acquired for, and supplied to, the Darwin Tree of Life Project. Further, the Wellcome Sanger Institute employs a process whereby due diligence is carried out proportionate to the nature of the materials themselves, and the circumstances under which they have been/are to be collected and provided for use. The purpose of

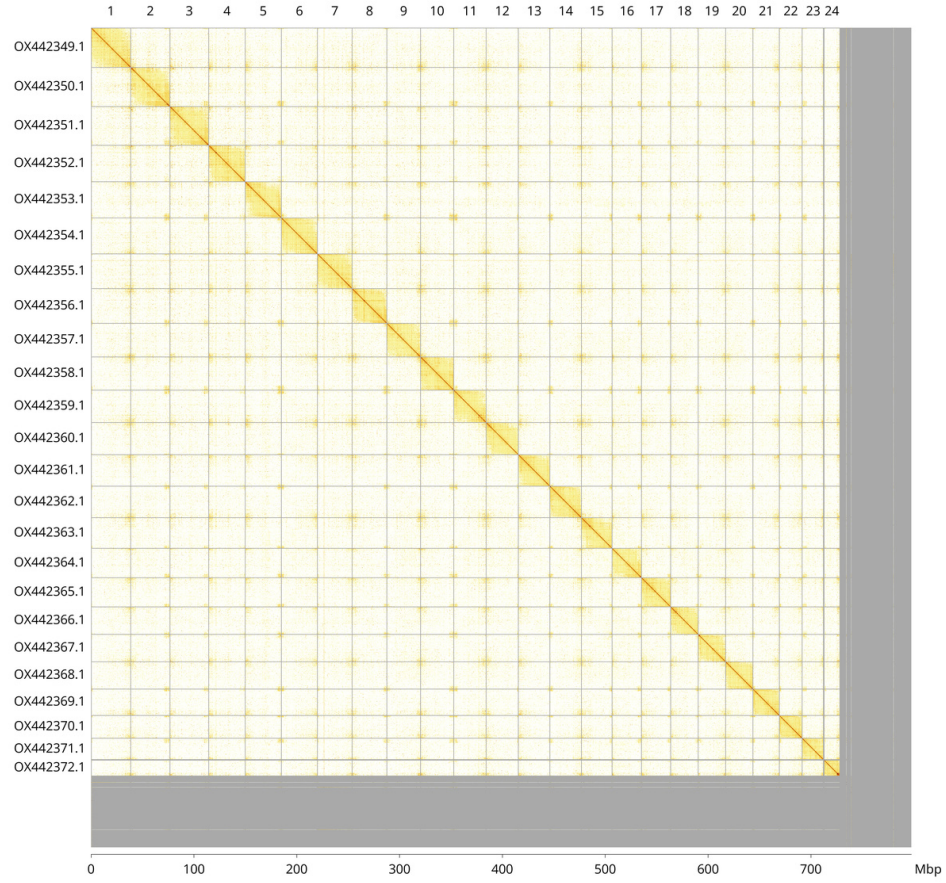

Figure 3. Hi-C contact map of the *Hyperoplus immaculatus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Hyperoplus immaculatus* fHypImm3.

INSDC accession	Molecule	Length (Mb)	GC%
OX442349.1	1	38.61	40.50
OX442350.1	2	38.12	41
OX442351.1	3	37.74	41
OX442352.1	4	35.48	41
OX442353.1	5	35.12	41
OX442354.1	6	35.09	41
OX442355.1	7	33.80	41
OX442356.1	8	33.75	41
OX442357.1	9	32.72	41
OX442358.1	10	32.33	41
OX442359.1	11	31.45	41
OX442360.1	12	31.41	41
OX442361.1	13	30.68	41
OX442362.1	14	30.56	41
OX442363.1	15	29.95	41
OX442364.1	16	28.54	41.50
OX442365.1	17	28.40	40.50
OX442366.1	18	26.76	41
OX442367.1	19	26.61	41
OX442368.1	20	26.59	41
OX442369.1	21	25.65	41
OX442370.1	22	22.33	41
OX442371.1	23	21.25	41
OX442372.1	24	15.26	41

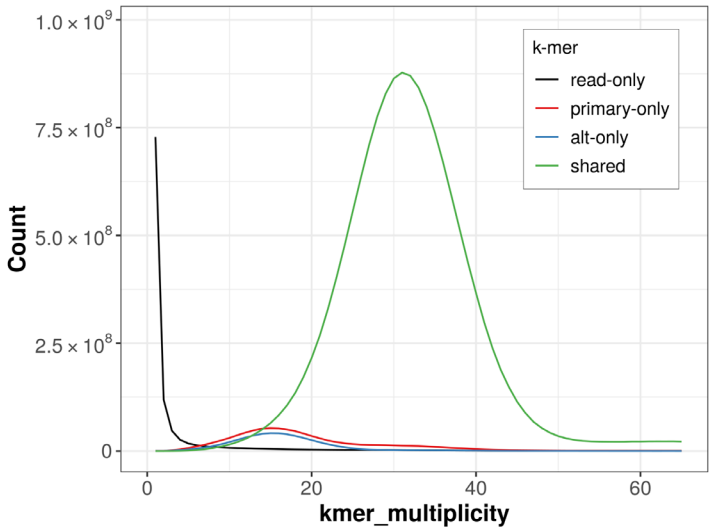

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers that appear in the reads but are not assembled. The green curve corresponds to *k*-mers shared by both haplotypes, and the red and blue curves show *k*-mers found only in one of the haplotypes.

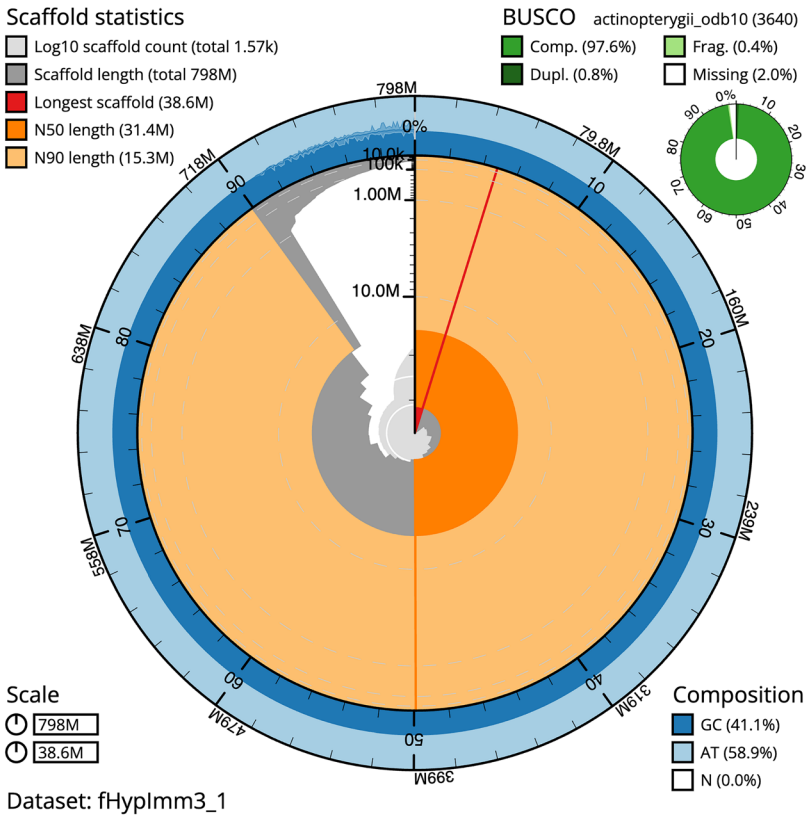

Figure 5. Assembly metrics for fHypImm3.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks display the distribution of GC, AT, and N percentages across the bins. Scaffolds are arranged clockwise from longest to shortest and are depicted in dark grey. The longest scaffold is indicated by the red arc, and the deeper orange and pale orange arcs represent the N50 and N90 lengths. A light grey spiral at the centre shows the cumulative scaffold count on a logarithmic scale. A summary of complete, fragmented, duplicated, and missing BUSCO genes in the set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

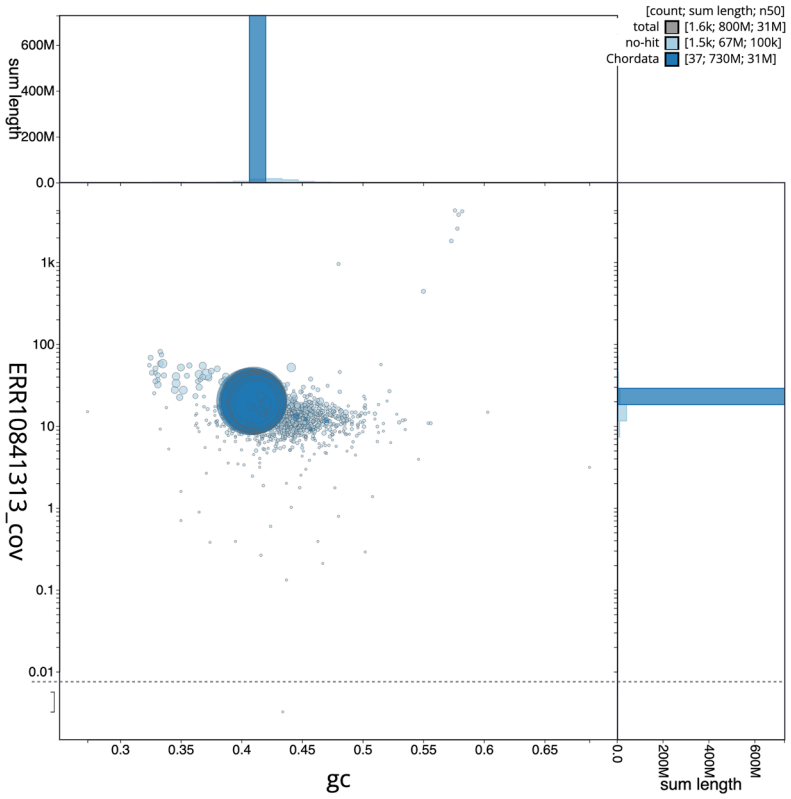

Figure 6. BlobToolKit GC-coverage plot for fHypImm3.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display the total length of sequences distributed across different levels of coverage and GC content. An interactive version of this figure is available on the BlobToolKit viewer.

Table 4. Earth Biogenome Project summary metrics for the *Hyperoplus immaculatus* assembly.

Measure	Value	Benchmark
EBP summary (primary)	6.C.Q53	6.C.Q40
Contig N50 length	1.49 Mb	≥ 1 Mb
Scaffold N50 length	31.41 Mb	= chromosome N50
Consensus quality (QV)	Primary: 53.1; alternate: 54.3; combined: 53.6	≥ 40
k-mer completeness	Primary: 93.04%; alternate: 90.14%; combined: 98.45%	≥ 95%
BUSCO	C:97.6%[S:96.8%;D:0.8%]; F:0.4%;M:2.0%;n:3640	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	91.28%	≥ 90%

this is to address and mitigate any potential legal and/or ethical implications of receipt and use of the materials as part of the research project, and to ensure that in doing so we align with best practice wherever possible. The overarching areas of consideration are:

- Ethical review of provenance and sourcing of the material
- Legality of collection, transfer and use (national and international)

Each transfer of samples is further undertaken according to a Research Collaboration Agreement or Material Transfer Agreement entered into by the Darwin Tree of Life Partner, Genome Research Limited (operating as the Wellcome Sanger Institute),

and in some circumstances, other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: Hyperoplus immaculatus (greater sand-eel). Accession number PRJEB58244. The genome sequence is released openly for reuse. The *Hyperoplus immaculatus* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), the Sanger Institute Tree of Life Programme (PRJEB43745) and the Vertebrate Genomes Project (PRJNA489243). All raw sequence data and the assembly have been deposited in INSDC databases. Raw data and assembly accession identifiers are reported in Table 1 and Table 2.

Production code used in genome assembly at the WSI Tree of Life is available at https://github.com/sanger-tol. Table 5 lists software versions used in this study.

Author information

Contributors are listed at the following links:

- Members of the Marine Biological Association Genome Acquisition Lab
- Members of the Darwin Tree of Life Barcoding collective
- Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team
- Members of Wellcome Sanger Institute Scientific Operations – Sequencing Operations
- Members of the Wellcome Sanger Institute Tree of Life Core Informatics team
- Members of the Tree of Life Core Informatics collective
- Members of the Darwin Tree of Life Consortium

Table 5. Software versions and sources.

Software	Version	Source
BEDTools	2.30.0	https://github.com/arq5x/bedtools2
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
fasta_windows	0.2.4	https://github.com/tolkit/fasta_windows
FastK	1.1	https://github.com/thegenemyers/FASTK
GenomeScope2.0	2.0.1	https://github.com/tbenavi1/genomescope2.0
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoaT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.16.1-r375	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
MitoHiFi	2	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.3	https://github.com/dfguan/purge_dups
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.4.0	https://github.com/sanger-tol/treeval
YaHS	1.2a	https://github.com/c-zhou/yahs

References

Allio R, Schomaker-Bastos A, Romiguier J, *et al.*: **MitoFinder: efficient automated large-scale extraction of mitogenomic data in target enrichment phylogenomics.** *Mol Ecol Resour.* 2020; **20**(4): 892–905.
PubMed Abstract | Publisher Full Text | Free Full Text

Altschul SF, Gish W, Miller W, *et al.*: **Basic Local Alignment Search Tool.** *J Mol Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

Bateman A, Martin MJ, Orchard S, *et al.*: **UniProt: the Universal Protein Knowledgebase in 2023.** *Nucleic Acids Res.* 2023; **51**(D1): D523–D531.
PubMed Abstract | Publisher Full Text | Free Full Text

Blaxter M, Mieszkowska N, Di Palma F, *et al.*: **Sequence locally, think globally: the Darwin Tree of Life project.** *Proc Natl Acad Sci U S A.* 2022; **119**(4): e2115642118.
PubMed Abstract | Publisher Full Text | Free Full Text

Buchfink B, Reuter K, Drost HG: **Sensitive protein alignments at Tree-of-Life scale using DIAMOND.** *Nat Methods.* 2021; **18**(4): 366–368.
PubMed Abstract | Publisher Full Text | Free Full Text

Challis R, Richards E, Rajan J, *et al.*: **BlobToolKit – interactive quality assessment of genome assemblies.** *G3 (Bethesda).* 2020; **10**(4): 1361–1374.
PubMed Abstract | Publisher Full Text | Free Full Text

Cheng H, Concepcion GT, Feng X, *et al.*: **Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm.** *Nat Methods.* 2021; **18**(2): 170–175.
PubMed Abstract | Publisher Full Text | Free Full Text

Corbin PG: ***Ammodytes immaculatus*, a new species of sand-eel found in European seas.** *Nature.* 1950; **166**(4221): 525–26.
PubMed Abstract | Publisher Full Text

Crowley L, Allen H, Barnes I, *et al.*: **A sampling strategy for genome sequencing the British terrestrial arthropod fauna [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2023; **8**: 123.
PubMed Abstract | Publisher Full Text | Free Full Text

Danecek P, Bonfield JK, Liddle J, *et al.*: **Twelve years of SAMtools and BCFtools.** *GigaScience.* 2021; **10**(2): giab008.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels P, Magnusson M, Lundin S, *et al.*: **MultiQC: summarize analysis results for multiple tools and samples in a single report.** *Bioinformatics.* 2016; **32**(19): 3047–3048.
PubMed Abstract | Publisher Full Text | Free Full Text

Ewels PA, Peltzer A, Fillinger S, *et al.*: **The nf-core framework for community-curated bioinformatics pipelines.** *Nat Biotechnol.* 2020; **38**(3): 276–278.
PubMed Abstract | Publisher Full Text

Formenti G, Abueg L, Brajuka A, *et al.*: **Gfastats: conversion, evaluation and manipulation of genome sequences using assembly graphs.** *Bioinformatics.* 2022; **38**(17): 4214–4216.
PubMed Abstract | Publisher Full Text | Free Full Text

Froese R, Pauly D: **FishBase: species summary for *Hyperoplus immaculatus*.** 2025.
Reference Source

Gomes P, Oliveira L, Oliveira R, *et al.*: **Range extension of *Hyperoplus immaculatus* (corbin, 1950) (Pisces, Ammodytidae) in Portuguese continental waters.** *J Appl Ichthyol.* 2018; **34**: 1331–34.
Publisher Full Text

Guan D, McCarthy SA, Wood J, *et al.*: **Identifying and removing haplotypic duplication in primary genome assemblies.** *Bioinformatics.* 2020; **36**(9): 2896–2898.
PubMed Abstract | Publisher Full Text | Free Full Text

Howard C, Denton A, Jackson B, *et al.*: **On the path to reference genomes for all biodiversity: lessons learned and laboratory protocols created in the Sanger Tree of Life core laboratory over the first 2000 species.** *bioRxiv.* 2025.
Publisher Full Text

Howe K, Chow W, Collins J, *et al.*: **Significantly improving the quality of genome assemblies through curation.** *GigaScience.* 2021; **10**(1): giaa153.
PubMed Abstract | Publisher Full Text | Free Full Text

Kerpedjiev P, Abdennur N, Lekschas F, *et al.*: **HiGlass: web-based visual exploration and analysis of genome interaction maps.** *Genome Biol.* 2018; **19**(1): 125.
PubMed Abstract | Publisher Full Text | Free Full Text

Kurtzer GM, Sochat V, Bauer MW: **Singularity: scientific containers for mobility of compute.** *PLoS One.* 2017; **12**(5): e0177459.
PubMed Abstract | Publisher Full Text | Free Full Text

Lawniczak MKN, Davey RP, Rajan J, *et al.*: **Specimen and sample metadata standards for biodiversity genomics: a proposal from the Darwin Tree of Life project [version 1; peer review: 2 approved with reservations].** *Wellcome Open Res.* 2022; **7**: 187.
Publisher Full Text

Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics.* 2018; **34**(18): 3094–3100.
PubMed Abstract | Publisher Full Text | Free Full Text

Manni M, Berkeley MR, Seppey M, *et al.*: **BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes.** *Mol Biol Evol.* 2021; **38**(10): 4647–4654.
PubMed Abstract | Publisher Full Text | Free Full Text

Merkel D: **Docker: lightweight Linux containers for consistent development and deployment.** *Linux J.* 2014; **2014**(239): 2.
Reference Source

O'Leary NA, Cox E, Holmes JB, *et al.*: **Exploring and retrieving sequence and metadata for species across the Tree of Life with NCBI Datasets.** *Sci Data.* 2024; **11**(1): 732.
PubMed Abstract | Publisher Full Text | Free Full Text

Ranallo-Benavidez TR, Jaron KS, Schatz MC: **GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes.** *Nat Commun.* 2020; **11**(1): 1432.
PubMed Abstract | Publisher Full Text | Free Full Text

Rao SSP, Huntley MH, Durand NC, *et al.*: **A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.** *Cell.* 2014; **159**(7): 1665–1680.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, McCarthy SA, Fedrigo O, *et al.*: **Towards complete and error-free genome assemblies of all vertebrate species.** *Nature.* 2021; **592**(7856): 737–746.
PubMed Abstract | Publisher Full Text | Free Full Text

Rhie A, Walenz BP, Koren S, *et al.*: **Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol.* 2020; **21**(1): 245.
PubMed Abstract | Publisher Full Text | Free Full Text

Schoch CL, Ciufo S, Domrachev M, *et al.*: **NCBI Taxonomy: a comprehensive update on curation, resources and tools.** *Database (Oxford).* 2020; **2020**: baaa062.
PubMed Abstract | Publisher Full Text | Free Full Text

Twyford AD, Beasley J, Barnes I, *et al.*: **A DNA barcoding framework for taxonomic verification in the Darwin Tree of Life project [version 1; peer review: 2 approved].** *Wellcome Open Res.* 2024; **9**: 339.
PubMed Abstract | Publisher Full Text | Free Full Text

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, *et al.*: **MitoHiFi: a python pipeline for mitochondrial genome assembly from PacBio high fidelity reads.** *BMC Bioinformatics.* 2023; **24**(1): 288.
PubMed Abstract | Publisher Full Text | Free Full Text

Vasimuddin M, Misra S, Li H, *et al.*: **Efficient architecture-aware acceleration of BWA-MEM for multicore systems.** In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS).* IEEE, 2019; 314–324.
Publisher Full Text

Zhou C, McCarthy SA, Durbin R: **YaHS: Yet another Hi-C Scaffolding tool.** *Bioinformatics.* 2023; **39**(1): btac808.
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ?

Version 1

Reviewer Report 29 December 2025

https://doi.org/10.21956/wellcomeopenres.27639.r140516

© **2025 Barros-García D.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

David Barros-García
Centro Interdisciplinar de Investigação Marinha E Ambiental (CIIMAR/CIMAR), Matosinhos, Portugal

The proposed article presents the first genome of the species Hyperoplus immaculatus (Corbin, 1950). To this end, the authors have used a set of techniques that includes PacBio reads and HiC libraries. The genome obtained is of normal size for a Hyperoplus immaculatus (Corbin, 1950) fish and has shown good quality indicators (97.6% BUSCO) and 92.28% of the genome assigned to 24 chromosome-level scaffolds.

I have no objections to the technical aspects of the manuscript, but I do feel that some lines comparing the results with similar genomes are missing, as well as the usefulness of this genome for other studies (conservation, evolution, biogeography, etc.).

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research focuses on integrated taxonomy, population genetics and evolutionary biology.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Reviewer Report 11 December 2025

https://doi.org/10.21956/wellcomeopenres.27639.r140515

© 2025 Zhang X. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Xizhao Zhang
Freshwater Fisheries and Germplasm Resources Utilization, Wuxi, China

This research presents a high-quality chromosome-level genome assembly for *Hyperoplus immaculatus*. The assembly utilizes PacBio HiFi sequencing reads, constructs scaffolds using Hi-C data, and undergoes annotation via RNA sequencing data. The primary assembly data covers 797.79 Mb of sequence, comprising 1,567 scaffolds. Of this, 91.28% of the sequence was assembled into 24 chromosome pseudomolecules, along with a 16.53 kb mitochondrial genome. This represents the second genome assembly for the genus *Hyperoplus* and will deepen our understanding of the sand eel's ecology—a species of significant research value as a key prey fish for marine predators.
Regarding two suggestions for this paper:
1: In the Genome assembly and Assembly quality assessment sections, the authors detail the software and pipeline used but lack descriptions of specific software parameters. If default parameters were used, please state this.
2: The genome report section lacks a description of the mitochondrial genome; it is recommended to add this.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Fisheries Resources，Genomics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 25 November 2025

https://doi.org/10.21956/wellcomeopenres.27639.r139621

© **2025 Jeffery N.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

?

Nick Jeffery
Bedford Institute of Oceanography, Fisheries and Oceans Canada, Canada, Canada

This article provides a summary of the greater sandeel genome, assembled using PacBio reads, HiC libraries, and RNA sequencing. The genome was assessed to be approximately 97% complete based on the estimated genome size and BUSCO scores. The paper is technically sound, but I do have the following questions:

Did the authors assemble a transcriptome from the RNA reads, or was the RNAseq just used for annotation through the Ensembl pipeline?

Could the authors provide a summary of coding vs non-coding proportions of the genome, including transposable element content? Software such as Earl Grey has made transposable element identification quite simple, and is useful information for genome evolution studies?

Does the number of pseduo-chromosomes match what is expected for this family/genus? It appears so, so the authors could cite a reference on this (e.g., Ocalewicz et al. 2019 Mar Biol Research)

A short paragraph or sentence at the end of the manuscript could help state the importance of this species. Are additional population genetics studies being conducted on this species or is there a specific conservation focus? I recognize the short nature of these data reports, but a short rationale on why this species was chosen for genome assembly would be useful.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: My research focuses on aquatic genomics, conservation biology, and evolutionary biology. I specifically focus on population genomics of marine organisms, and genome assembly and evolution. My research also uses environmental DNA metabarcoding for biodiversity monitoring in the marine realm.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.