DATA NOTE

The genome sequence of the highfin lizardfish, *Bathysaurus mollis* Günther, 1878 (Aulopiformes: Bathysauridae)

[version 1; peer review: 4 approved]

Tammy Horton[1], Andrew R. Gates[1], Chris Fletcher[2],
Natural History Museum 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]National Oceanography Centre, Southampton, England, UK
[2]Natural History Museum, London, England, UK

v1 **First published:** 24 Oct 2025, **10**:605
https://doi.org/10.12688/wellcomeopenres.25077.1
Latest published: 24 Oct 2025, **10**:605
https://doi.org/10.12688/wellcomeopenres.25077.1

Abstract

We present a genome assembly from an individual *Bathysaurus mollis* (highfin lizardfish; Chordata; Actinopteri; Aulopiformes; Bathysauridae). The genome sequence has a total length of 1 065.77 megabases. Most of the assembly (95.16%) is scaffolded into 24 chromosomal pseudomolecules. The mitochondrial genome has also been assembled, with a length of 16.68 kilobases. 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

Bathysaurus mollis; highfin lizardfish; genome sequence; chromosomal; Aulopiformes

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓ ✓ ✓

	1	2	3	4
version 1 24 Oct 2025	✓ view	✓ view	✓ view	✓ view

1. **David Ray**, Texas Tech University, Lubbock, USA
2. **Yichen Dai**, Fudan University, Shanghai, China
3. **Anthony K. Redmond**, University College Dublin, Dublin, Ireland
4. **Salvatore D'Aniello**, Stazione Zoologica Anton Dohrn, Napoli, Italy
Anna Albanese, Universita degli Studi di Ferrara Dipartimento di Scienze della Vita e Biotecnologie, Ferrara, Italy

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: Horton T: Investigation, Resources, Writing – Original Draft Preparation, Writing – Review & Editing; **Gates AR**: Resources; **Fletcher C**: 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 Porcupine Abyssal Plain – Sustained Observatory of the Natural Environment Research Council (NERC, UK) was previously funded through the Climate Linked Atlantic Sector Science (CLASS) project supported by NERC National Capability funding (NE/R015953/1) and now through the AtlantiS program (NE/Y005589/1).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Horton T, Gates AR, Fletcher C *et al.* **The genome sequence of the highfin lizardfish, *Bathysaurus mollis* Günther, 1878 (Aulopiformes: Bathysauridae) [version 1; peer review: 4 approved]** Wellcome Open Research 2025, **10**:605 https://doi.org/10.12688/wellcomeopenres.25077.1

First published: 24 Oct 2025, **10**:605 https://doi.org/10.12688/wellcomeopenres.25077.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Actinopterygii; Actinopteri; Neopterygii; Teleostei; Osteoglossocephalai; Clupeocephala; Euteleosteomorpha; Neoteleostei; Eurypterygia; Aulopa; Aulopiformes; Alepisauroidei; Bathysauridae; *Bathysaurus*; *Bathysaurus mollis* Günther, 1878 (NCBI:txid1126214)

Background

We present a chromosome-level genome sequence for *Bathysaurus mollis* Günther (1878), the highfin lizardfish. The assembly was produced using the Tree of Life pipeline from a specimen collected from the Porcupine Abyssal Plain Sustained Observatory in the Northeast Atlantic (Figure 1). This assembly 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).

There are two valid species of the genus *Bathysaurus*, *B. ferox* Günther, 1878 and *B. mollis*, both of which are bathydemersal species found in deep waters circumglobally at tropical and temperate latitudes. *Bathysaurus ferox* is known from depths of 1 000 to 2 500 m, while *B. mollis* is generally known from 2 500 to at least 4 850 m (Sulak *et al.*, 1985). Both species have been shown to be synchronous hermaphrodites and are predominantly piscivorous (Sulak *et al.*, 1985). There are currently no known threats to *B. mollis* and because of the depth distribution of this species, it is unlikely to be threatened by anthropogenic disturbances and is considered by the IUCN Red list to be of Least Concern (de Bruyne *et al.*, 2015).

Figure 1. Photographs of the *Bathysaurus mollis* Günther, 1878 specimen (DISCOLL_JC231_082_023; fBatMol1) used for genome sequencing. Top: lateral view of whole specimen on board prior to tissue sampling and preservation. Bottom: Dorsal detail view of head of the same specimen.

The two species can be distinguished by their colouration (*Bathysaurus mollis* is unpigmented with pale translucent flesh, while *B. ferox* is generally pigmented and darker grey-brown), by their dorsal fins (*B. mollis* has the dorsal fin base shorter than its head, whereas in *B. ferox* the dorsal fin base is much longer than its head); and by the possession of an adipose fin in *B. mollis* (lacking in *B. ferox*) (Sulak *et al.*, 1985).

Another chromosome-level assembly for this species is also available (GCA_048564825.1; submitted by IDSSE) (NCBI datasets, O'Leary *et al.*, 2024).

Methods

Sample acquisition and DNA barcoding

The specimen used for genome and RNA sequencing was an adult *Bathysaurus mollis* Günther, 1878. The sample was collected during RRS *James Cook* cruise, JC231, using an OTSB14 (semi-balloon otter trawl, 14 m headrope; Merritt & Marshall, 1981) from the Porcupine Abyssal Plain Sustained Observatory (PAP-SO) site located in the NE Atlantic in international waters (48° 53.176' N, 16° 27.503' W to 48° 53.151' N, 16° 36.704' W, at 4840–4844 m depth, on 13th May 2022). The PAP-SO site has been the focus of an open ocean and deep-seabed study programme since 1985 (Hartman *et al.*, 2021; Hartman, 2022). Once on board the trawl sample was washed in filtered sea water, and selected specimens were photographed and tissue samples removed prior to fixation in 95% ethanol or 4% buffered formaldehyde. Part of the dissected tissue sample was placed into 95% ethanol for barcoding, and the rest of the tissue was placed into a 0.7 ml cryovial and preserved without fixative at –85 °C for whole genome sequencing. The specimen was sampled by Chris Fletcher and was identified by Tammy Horton based on morphology. Voucher tissue material is lodged at the Natural History Museum in London (NHM registration number NHMUK014453704; ToLID fBatMol1), while the voucher specimen was initially preserved in 4% buffered formaldehyde and transferred to 80% ethanol and is retained in the Discovery Collections, National Oceanography Centre, Southampton with specimen number DISCOLL_JC231_082_023. For the Darwin Tree of Life sampling and metadata approach, refer to 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 fBatMol1 sample was weighed and triaged to determine the appropriate extraction protocol. Muscle tissue was homogenised by powermashing using a PowerMasher II tissue disruptor. 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 2.84 ng/μL and a yield of 369.20 ng, with a fragment size of kb. The 260/280 spectrophotometric ratio was 2.98, and the 260/230 ratio was 6.41.

RNA was extracted from muscle tissue of fBatMol1 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 on a Revio instrument (Pacific Biosciences). The prepared library was normalised to 2 nM, and 15 μL was used for making complexes. Primers were annealed and polymerases bound to generate circularised complexes, following the manufacturer's instructions. Complexes were purified using 1.2X SMRTbell beads, then diluted to the Revio loading concentration (200–300 pM) and spiked with a Revio sequencing internal control. The sample was sequenced on a Revio 25M SMRT cell. The SMRT Link software (Pacific Biosciences), a web-based workflow manager, was used to configure and monitor the run and to carry out primary and secondary data analysis.

Hi-C

Sample preparation and crosslinking

The Hi-C sample was prepared from 20–50 mg of frozen muscle tissue of the fBatMol1 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 to 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. 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 X.

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 X 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. 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 53 breaks and 150 joins. This reduced the scaffold count by 6.6% and increased scaffold N50 by 1.7%. 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 *Bathysaurus mollis* specimen generated 27.55 Gb (gigabases) from 2.56 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 796.29 Mb, with a heterozygosity of 1.61% and repeat content of 31.45% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 32× coverage. Hi-C sequencing produced 88.96 Gb from 589.11 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 1 065.77 Mb in 1 095 scaffolds, with 4 638 gaps, and a scaffold N50 of 44.26 Mb (Table 2).

Most of the assembly sequence (95.16%) 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 (length 16.68 kb, OZ257237.1). 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 52.2. The *k*-mer completeness is 82.86% for the primary assembly, 78.18% for the alternate haplotype, and 98.50% for the combined assemblies (Figure 4).

BUSCO v.5.7.1 analysis using the actinopterygii_odb10 reference set (n = 3 640) identified 96.8% of the expected gene set (single = 95.5%, duplicated = 1.3%). 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 **5.C.Q50**.

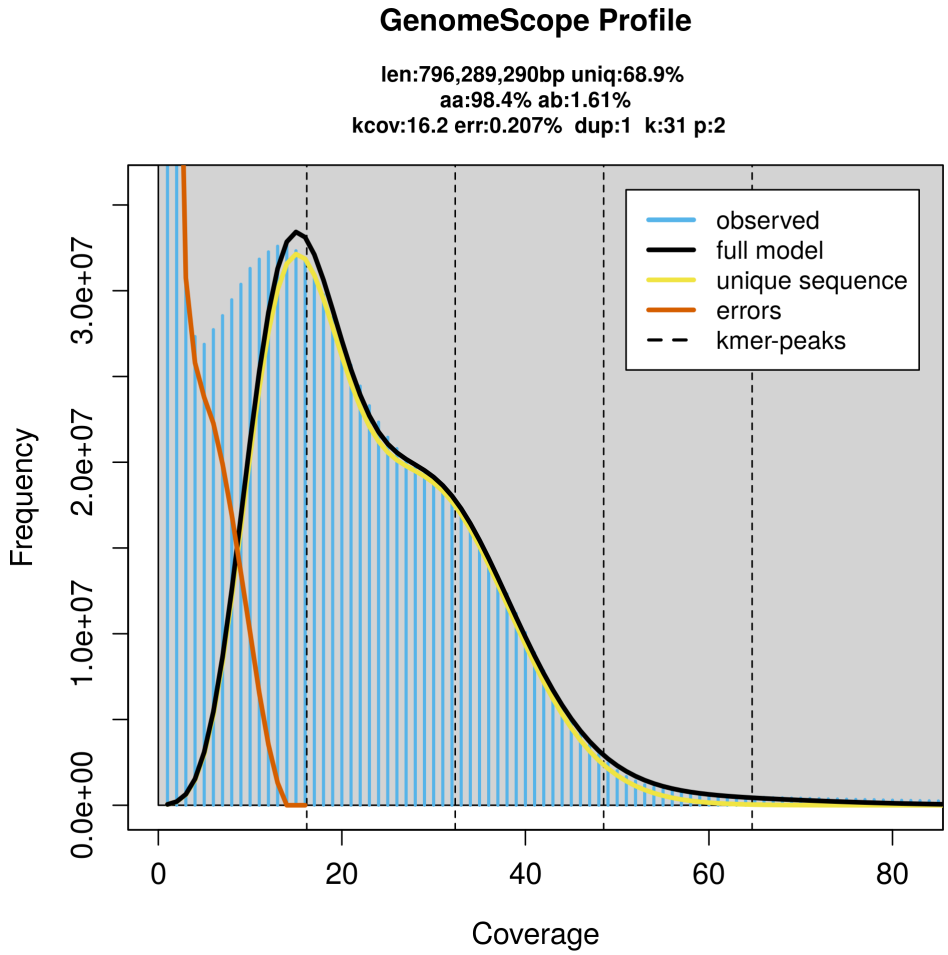

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB76377.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	fBatMol1	fBatMol1	fBatMol1
Specimen ID	NHMUK014453704	NHMUK014453704	NHMUK014453704
BioSample (source individual)	SAMEA114806302	SAMEA114806302	SAMEA114806302
BioSample (tissue)	SAMEA114806440	SAMEA114806440	SAMEA114806441
Tissue	muscle	muscle	muscle
Instrument	Revio	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR13245303	ERR13248980	ERR14379122
Read count total	2.56 million	589.11 million	74.82 million
Base count total	27.55 Gb	88.96 Gb	11.30 Gb

Table 2. Genome assembly statistics.

Assembly name	fBatMol1.1
Assembly accession	GCA_965279225.1
Alternate haplotype accession	GCA_965279235.1
Assembly level	chromosome
Span (Mb)	1 065.77
Number of chromosomes	24
Number of contigs	5 733
Contig N50	0.41 Mb
Number of scaffolds	1 095
Scaffold N50	44.26 Mb
Organelles	Mitochondrion: 16.68 kb

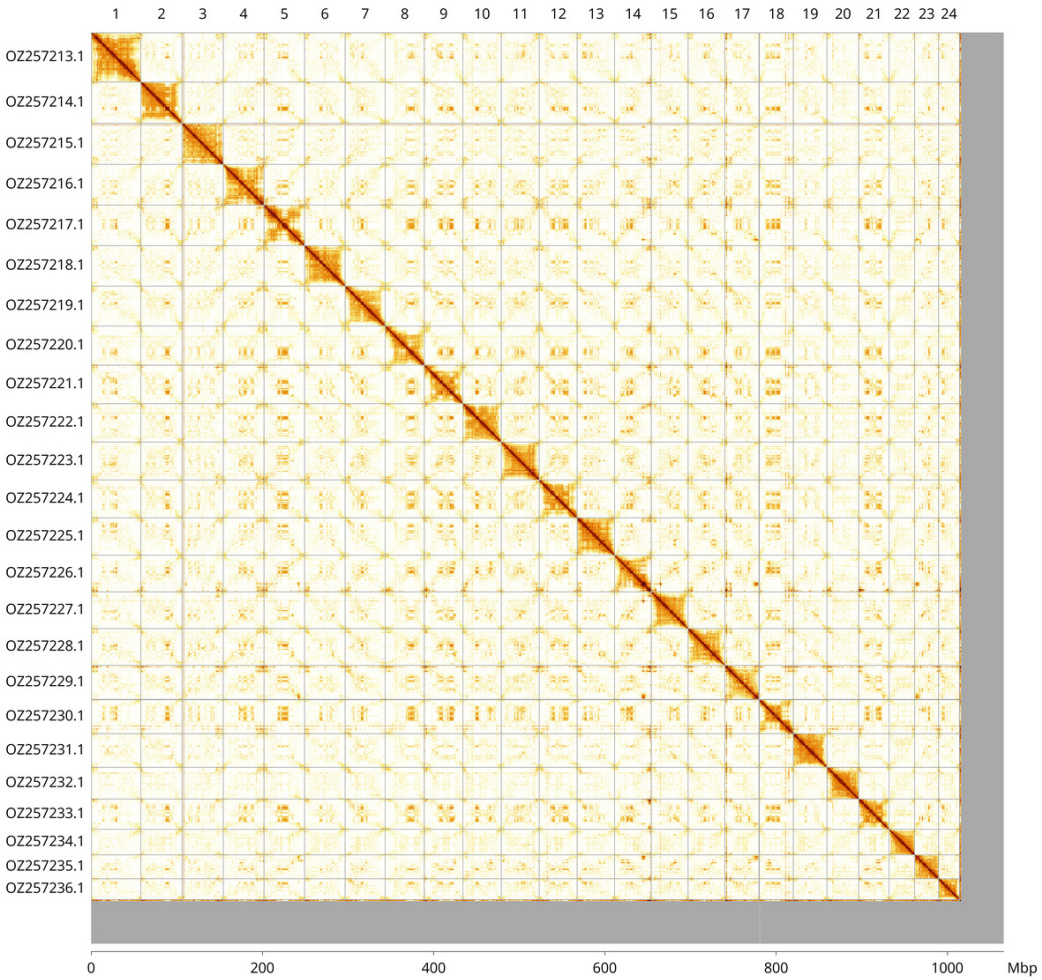

Figure 3. Hi-C contact map of the *Bathysaurus mollis* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the primary genome assembly of *Bathysaurus mollis* fBatMol1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ257213.1	1	57.87	42
OZ257214.1	2	48.43	42
OZ257215.1	3	47.95	42
OZ257216.1	4	47.73	42
OZ257217.1	5	47.39	41.50
OZ257218.1	6	47.35	42
OZ257219.1	7	46.57	42.50
OZ257220.1	8	45.69	42
OZ257221.1	9	45.12	42.50
OZ257222.1	10	44.92	42
OZ257223.1	11	44.45	42
OZ257224.1	12	44.26	42
OZ257225.1	13	43.56	41.50
OZ257226.1	14	43.09	42.50
OZ257227.1	15	42.90	42
OZ257228.1	16	42.75	42
OZ257229.1	17	40.41	42.50
OZ257230.1	18	39.78	42.50
OZ257231.1	19	39.28	42
OZ257232.1	20	36.96	42.50
OZ257233.1	21	35.59	42.50
OZ257234.1	22	29.91	42.50
OZ257235.1	23	27.71	42
OZ257236.1	24	24.49	42.50

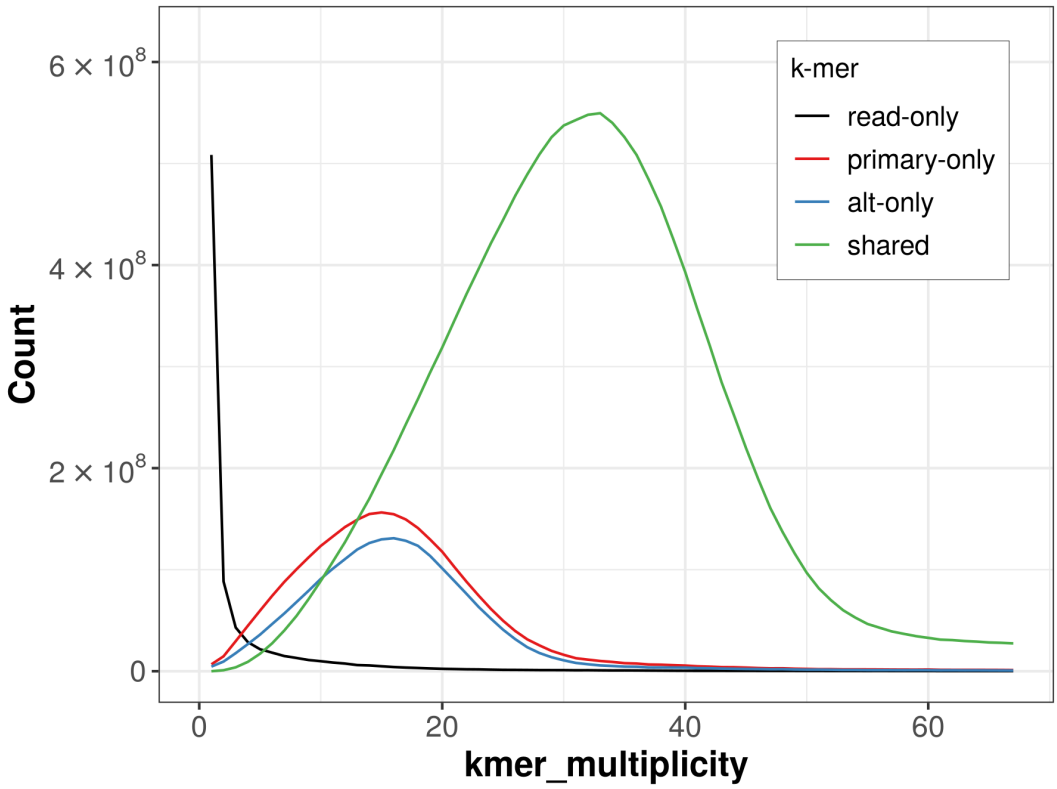

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers 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.

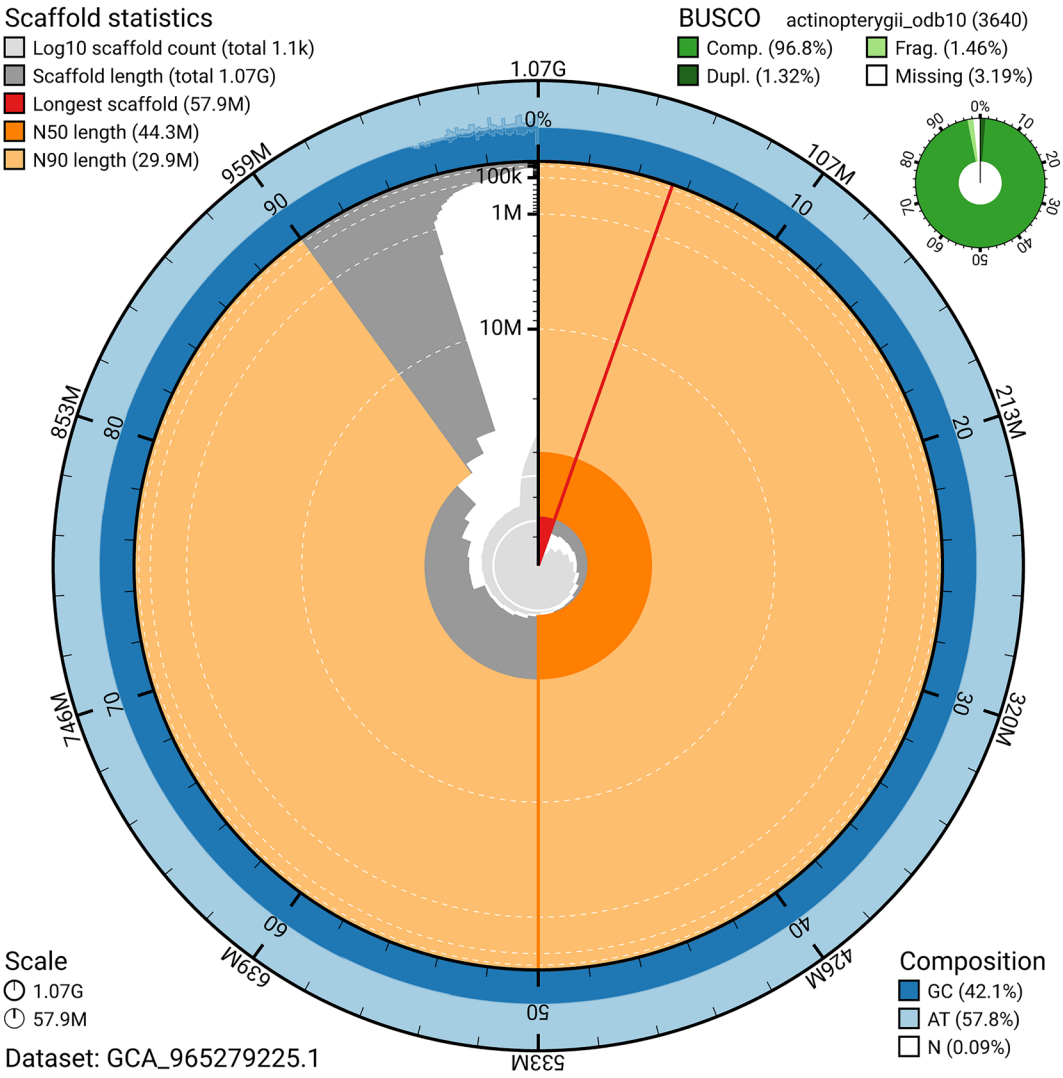

Figure 5. Assembly metrics for fBatMol1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue tracks 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 actinopterygii_odb10 set is presented at the top right. An interactive version of this figure can be accessed on the BlobToolKit viewer.

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 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)

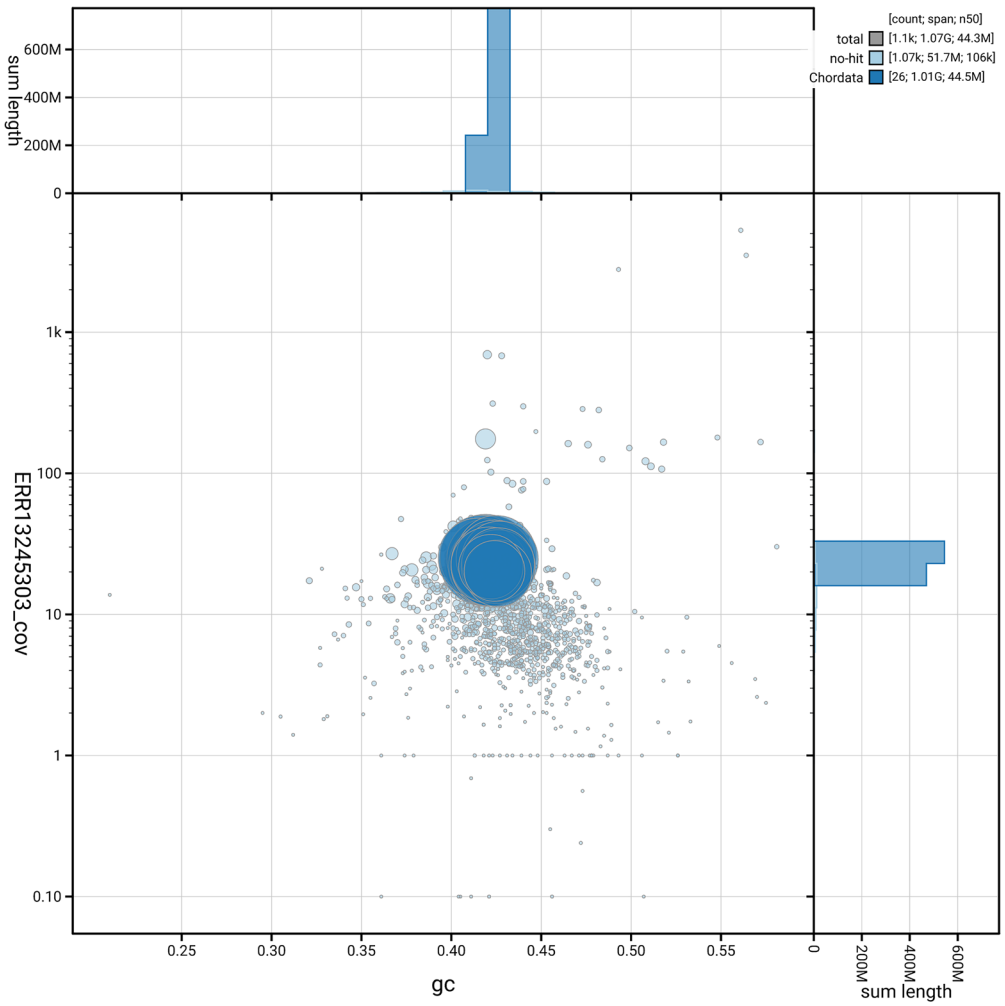

Figure 6. BlobToolKit GC-coverage plot for fBatMol1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes display 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 *Bathysaurus mollis* assembly.

Measure	Value	Benchmark
EBP summary (primary)	5.C.Q50	6.C.Q40
Contig N50 length	0.41 Mb	≥ 1 Mb
Scaffold N50 length	44.26 Mb	= chromosome N50
Consensus quality (QV)	Primary: 50.9; alternate: 53.1; combined: 52.2	≥ 40
k-mer completeness	Primary: 82.86%; alternate: 78.18%; combined: 98.50%	≥ 95%
BUSCO	C:96.8% [S:95.5%; D:1.3%]; F:1.5%; M:1.7%; n:3 640	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	95.16%	≥ 90%

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: Bathysaurus mollis. Accession number PRJEB76377. The genome sequence is released openly for reuse. The *Bathysaurus mollis* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), the Sanger Institute Tree of Life Programme (PRJEB43745) and Vertebrate Genomes Project (PRJNA489243). All raw sequence data and the assembly have been deposited in INSDC databases. The genome will be annotated using available RNA-Seq data and presented through the Ensembl pipeline at the European Bioinformatics Institute. 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.

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+/
BlobToolKit	4.4.5	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.7.1	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
Cooler	0.8.11	https://github.com/open2c/cooler
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
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	0.19.8-r603	https://github.com/chhylp123/hifiasm
HiGlass	1.13.4	https://github.com/higlass/higlass
MerquryFK	1.1.2	https://github.com/thegenemyers/MERQURY.FK
Minimap2	2.28-r1209	https://github.com/lh3/minimap2
MitoHiFi	3	https://github.com/marcelauliano/MitoHiFi
MultiQC	1.14; 1.17 and 1.18	https://github.com/MultiQC/MultiQC
Nextflow	24.10.4	https://github.com/nextflow-io/nextflow
PretextSnapshot	-	https://github.com/sanger-tol/PretextSnapshot
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
samtools	1.21	https://github.com/samtools/samtools
sanger-tol/ascc	0.1.0	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	v0.7.1	https://github.com/sanger-tol/blobtoolkit
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.2	https://github.com/c-zhou/yahs

Author information

Contributors are listed at the following links:

- Members of the Natural History Museum 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

Acknowledgements

The specimen collection would not be possible without the ongoing support of the National Marine Facilities group, the Ocean Technology and Engineering group, and the Marine Autonomous Robotic Systems groups at the National Oceanography Centre. We thank the captain and crew of the RRS James Cook cruise JC231 involved in the sample collection, including equipment deployment and recovery. We are grateful to the scientists on the benthic team who dealt with the specimens at sea, and in particular to Brian Bett for his dedication to the OTSB trawl fishing.

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

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

de Bruyne G, Carpenter KE, Smith-Vaniz WF: ***Bathysaurus mollis***. The IUCN *Red List of Threatened Species*, 2015.
Publisher 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

Günther A: **Preliminary notices of deep-sea fishes collected during the voyage of H.M.S. Challenger.** *Ann mag nat hist.* 1878; **2**: 17–28.
Publisher Full Text

Hartman SE: **RRS James Cook Cruise 231, 01 May - 19 May 2022: Time-series studies at the Porcupine Abyssal Plain Sustained Observatory**. 2022. **201**.
Reference Source

Hartman SE, Bett BJ, Durden JM, *et al.*: **Enduring science: Three decades of observing the Northeast Atlantic from the Porcupine Abyssal Plain Sustained Observatory (PAPSO).** *Prog Oceanogr.* 2021; **191**: 102508.
Publisher 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**

Sulak KJ, Wenner CA, Sedberry GR, *et al.*: **The life history and systematics of deep-sea lizard fishes, genus *Bathysaurus* (Synodontidae).** *Can J Zool.* 1985; **63**(3): 623–42.
Publisher 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 27 November 2025

https://doi.org/10.21956/wellcomeopenres.27644.r138505

✔ **Salvatore D'Aniello**
Stazione Zoologica Anton Dohrn, Napoli, Italy
Anna Albanese
Universita degli Studi di Ferrara Dipartimento di Scienze della Vita e Biotecnologie, Ferrara, Emilia-Romagna, Italy

The present manuscript regards the genome sequencing of highfin lizardfish *Bathysaurus mollis,* whitin the Darwin Tree of Life iniciative.

We believe that the quality of the experimental protocols and computational analyses are highly standardized and in general of high quality.

The present project will provide material for comparative genomics studies, and therefore we support its publication.

We would ask to the authors to revise the spectrophotometric values of nucleic acid extraction, since they declare very high values: the 260/280 ratio was 2.98, and the 260/230 ratio was 6.41.

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: Evolutionary and Developmental Biology (EvoDevo)

We confirm that we have read this submission and believe that we 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.27644.r138504

Anthony K. Redmond
University College Dublin, Dublin, Ireland

The study encompassing production of a genome assembly and raw RNA-seq data for the highfin lizardfish. The study appears to be performed to a high standard and data appropriately shared openly. The genome assembly appears to be of good accuracy and completeness. I am overall please with the study and have no major comments.

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: genome evolution

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 22 November 2025

https://doi.org/10.21956/wellcomeopenres.27644.r138508

Yichen Dai
Fudan University, Shanghai, Shanghai, China

The Darwin Tree of Life project is extremely important for advancing our understanding of evolution from a molecular perspective. The sampling and assembly of different eukaryotic species is a huge effort, not to also mention that the quality of these assembled genomes is extremely high. Specifically for this paper, the authors present a genome assembly for *Bathysaurus mollis*, which is an interesting resource for those investigating evolution of the fish lineage. I have no issues with the manuscript, and the methods and genome assembly described in the manuscript are detailed. In addition, I have verified that the genome data submitted are openly accessible. My only suggestion for the authors is that it would be very helpful to the readers if they could include photos or a figure when they describe the difference between *Bathysaurus mollis* and *B. ferox*.

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: Genome evolution, molecular evolution, evo-devo

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 12 November 2025

https://doi.org/10.21956/wellcomeopenres.27644.r138509

David Ray
Texas Tech University, Lubbock, Texas, USA

This is a standard data note for the Darwin Tree of Life project. All of the typical information is present and clearly presented.

I note only one small problem. Near the end of the Nucleic acid extraction paragraph, a number is missing before 'kb'. This should indicate the mean fragment size.

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: Vertebrate genonimcs and evolution

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.