DATA NOTE

The genome sequence of True's beaked whale, *Mesoplodon mirus* True, 1913 (Artiodactyla: Ziphiidae)

[version 1; peer review: 2 approved]

Nicholas J. Davison[1], Georg Hantke[2], Phillip A. Morin[3],
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]University of Glasgow, Glasgow, Scotland, UK
[2]National Museums Scotland, Edinburgh, Scotland, UK
[3]Southwest Fisheries Science Center, NOAA, La Jolla, California, USA

v1 **First published:** 06 Jan 2026, **11**:10
https://doi.org/10.12688/wellcomeopenres.25413.1
Latest published: 06 Jan 2026, **11**:10
https://doi.org/10.12688/wellcomeopenres.25413.1

Abstract

We present a genome assembly from an individual female *Mesoplodon mirus* (True's beaked whale; Chordata; Mammalia; Artiodactyla; Ziphiidae). The assembly contains two haplotypes with total lengths of 3 442.40 megabases and 2 956.53 megabases. Most of haplotype 1 (79.62%) is scaffolded into 21 chromosomal pseudomolecules, including the X sex chromosome. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 16.35 kilobases. Gene annotation of this assembly on Ensembl identified 18 422 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

Mesoplodon mirus; True's beaked whale; genome sequence; chromosomal; Artiodactyla

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 06 Jan 2026	✓ view	✓ view

1. **Jayan Duminda M Senevirathna**, Uva Wellassa University, Badulla, Sri Lanka

2. **Mariana F Nery**, State University of Campinas, São Paulo, Brazil

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: Davison NJ: Investigation, Resources, Writing – Review & Editing; **Hantke G**: Investigation, Resources; **Morin PA**: Writing – Original Draft Preparation, Writing – Review & Editing;

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.

How to cite this article: Davison NJ, Hantke G, Morin PA *et al.* **The genome sequence of True's beaked whale, *Mesoplodon mirus* True, 1913 (Artiodactyla: Ziphiidae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:10 https://doi.org/10.12688/wellcomeopenres.25413.1

First published: 06 Jan 2026, **11**:10 https://doi.org/10.12688/wellcomeopenres.25413.1

Species taxonomy

Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Laurasiatheria; Artiodactyla; Whippomorpha; Cetacea; Odontoceti; Ziphiidae; *Mesoplodon*; *Mesoplodon mirus* True, 1913 (NCBI:txid52113)

Background

The genus *Mesoplodon* is the most speciose genus of the cetacean infraorder, with 16 recognised species (Committee on Taxonomy & Society for Marine Mammalogy, 2025). True's whale (*Mesoplodon mirus*) previously included two disjunct antitropical types, but was recently divided into two species in the North Atlantic (*M. mirus*) and Southern Ocean (*M. eueu*) based on morphological and genomic data (Carroll *et al.*, 2021). Like other Mesoplodont species, True's beaked whale is small- to medium-sized (≤ 5.3 m) and characterised by a single tooth in the bottom jaw projecting outside of the mouth (males only) used for intraspecific fighting among males, most likely related to establish breeding hierarchies (Pitman, 2018).

Few beaked whale species have been studied extensively, and many are virtually unknown, having been identified only from beach-stranded animals. They are difficult to observe due to typically far-offshore habitat, small group size (~1 to 6), deep diving behavior that limits time at the surface, and minimal surface activity. In general, Mesoplodont whales inhabit deep ocean (>2000 m) and continental slope regions, feeding primarily on small fish and squids. They use suction feeding to capture prey and swallow them whole (Pitman, 2018). True's beaked whales are rarely seen, and what's known of their diet comes primarily from stomach contents of a handful of stranded animals, indicating a diet of mostly small fish (at least for these few animals prior to stranding) (Hernandez-Milian *et al.*, 2017).

True's beaked whales are listed as Least Concern on the IUCN Red List of Threatened Species (Pitman *et al.*, 2022), but there is no information on global abundance. They are considered common in part of their range, but unknown in other parts. Threats include fishing and harvesting of aquatic resources, pollution, and habitat shifting and alteration (Pitman *et al.*, 2022).

The genome of True's beaked whale, *Mesoplodon mirus*, was sequenced as part of the Darwin Tree of Life Project, a collaborative effort to sequence all named eukaryotic species in the Atlantic Archipelago of Britain and Ireland. Here we present a chromosomal-level genome sequence for *Mesoplodon mirus* from the eastern North Atlantic, based on a specimen from the Isle of Lewis, Scotland (Figure 1).

Methods

Sample acquisition

The specimen used for genome sequencing was an adult female *Mesoplodon mirus* (specimen ID SAN00003683, ToLID mMesMir1; Figure 1), collected from Brager, Isle of Lewis, Western Isles, Scotland, United Kingdom (latitude 58.3474, longitude –6.6443) on 2023-11-22. The specimen was collected and identified by Nick Davison (Scottish Marine Animal Stranding Scheme University of Glasgow). 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 specimen is now deposited and registered at National Museums Scotland, reference number NMS.Z.2020.24.

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 mMesMir1 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the lung 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 manual 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 7.5 ng/μL and a yield of 975.00 ng. The 260/280 spectrophotometric ratio was 2.33, and the 260/230 ratio was 2.41.

RNA was extracted from lung tissue of mMesMir1 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

Figure 1. Photograph of the *Mesoplodon mirus* (mMesMir1) carcass from which samples were taken for genome sequencing (photo credit Nick Davison).

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 lung tissue from the mMesMir1 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 to create 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 in Hi-C phasing mode (Cheng *et al.*, 2021; Cheng *et al.*, 2022), producing two haplotypes. Hi-C reads (Rao *et al.*, 2014) were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). Contigs were further scaffolded with Hi-C data in YaHS (Zhou *et al.*, 2023), using 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 78 breaks and 318 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 both 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 *Mesoplodon mirus* specimen generated 172.02 Gb (gigabases) from 16.87 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 3 139.70 Mb, with a heterozygosity of 0.15% and repeat content of 35.59% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 53× coverage. Hi-C sequencing produced 300.62 Gb from 1 990.85 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 genome was assembled into two haplotypes using Hi-C phasing. Haplotype 1 was curated to chromosome level, while haplotype 2 was assembled to scaffold level. The final assembly

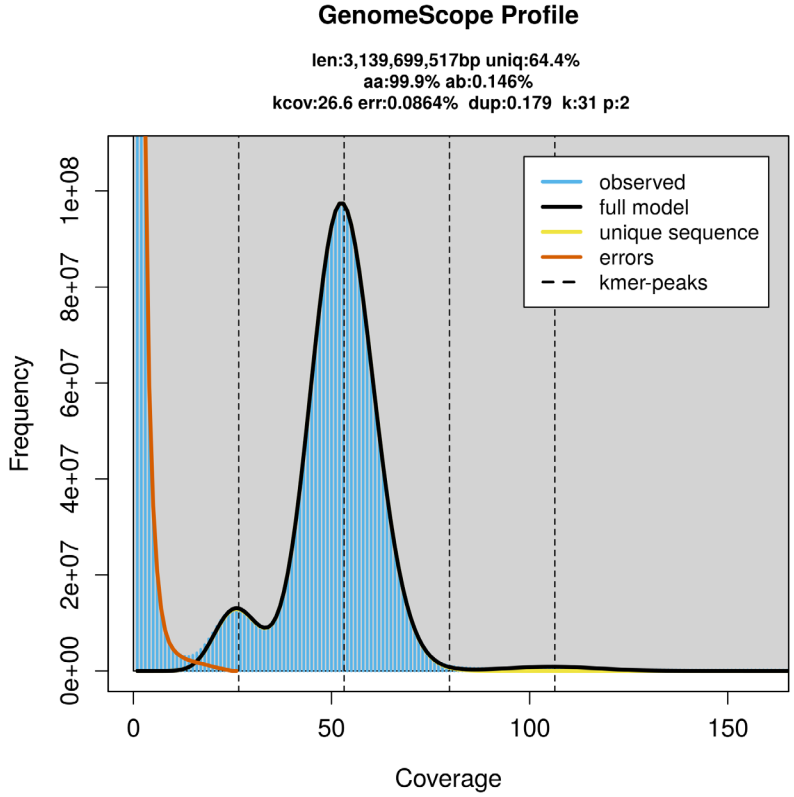

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

has a total length of 3 442.40 Mb in 2 756 scaffolds, with 1 240 gaps, and a scaffold N50 of 115.72 Mb (Table 2). Most of the assembly sequence (79.62%) was assigned to 21 chromosomal-level scaffolds, representing 20 autosomes and the X sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to synteny (Figure 3; Table 3). The X chromosome was identified through synteny analysis with the genome of *Globicephala melas* (GCA_963455315.2) (Davison *et al.*, 2025). During curation, we observed that the exact order and orientation of the contigs are unknown in the following regions:

- chromosome 1 (91 500–208 500 kbp)
- chromosome 3 (135 300–141 000 kbp)
- chromosome 4 (83 400–91 000 kbp)
- chromosome 7 (74 500–104 500 kbp)
- chromosome 10 (98 900–111 400 kbp)
- chromosome 18 (28 000–37 500 kbp).

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.

Table 1. Specimen and sequencing data for BioProject PRJEB79281.

Platform	PacBio HiFi	Hi-C	RNA-seq
ToLID	mMesMir1	mMesMir1	mMesMir1
Specimen ID	SAN00003683	SAN00003683	SAN00003683
BioSample (source individual)	SAMEA115358965	SAMEA115358965	SAMEA115358965
BioSample (tissue)	SAMEA115358966	SAMEA115358966	SAMEA115358966
Tissue	lung	lung	lung
Instrument	Revio	Illumina NovaSeq X	Illumina NovaSeq X
Run accessions	ERR13605520; ERR13605519	ERR13602182	ERR14379134
Read count total	16.87 million	1 990.85 million	70.02 million
Base count total	172.02 Gb	300.62 Gb	10.57 Gb

Table 2. Genome assembly statistics.

Assembly name	mMesMir1.hap1.1	mMesMir1.hap2.1
Assembly accession	GCA_964341445.1	GCA_964341485.1
Assembly level	chromosome	scaffold
Span (Mb)	3 442.40	2 956.53
Number of chromosomes	21	scaffold-level
Number of contigs	3 996	2 457
Contig N50	3.3 Mb	3.51 Mb
Number of scaffolds	2 756	1 384
Scaffold N50	115.72 Mb	107.9 Mb
Longest scaffold length (Mb)	232.78	-
Sex chromosomes	X	-
Organelles	Mitochondrion: 16.35 kb	-

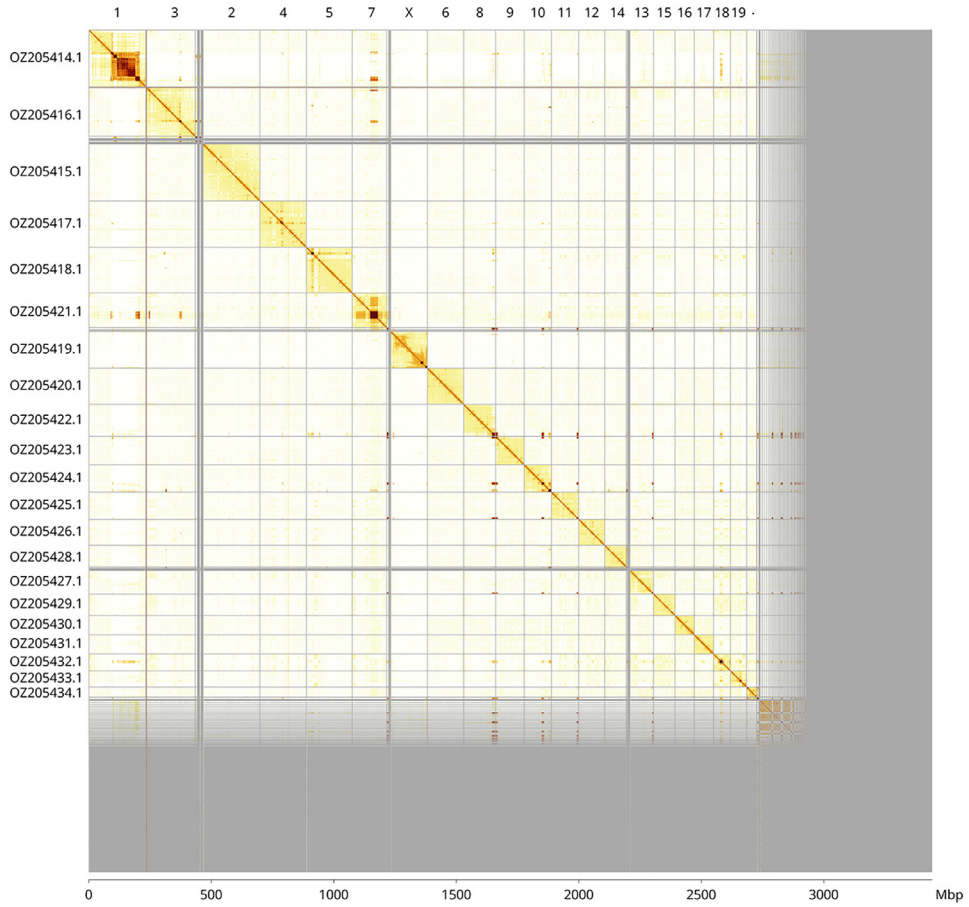

Figure 3. Hi-C contact map of the *Mesoplodon mirus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Mesoplodon mirus* mMesMir1.

INSDC accession	Molecule	Length (Mb)	GC%
OZ205414.1	1	234.69	43
OZ205415.1	2	231.52	40
OZ205416.1	3	232.89	42
OZ205417.1	4	189.20	42
OZ205418.1	5	187.08	41
OZ205420.1	6	147.74	39.50
OZ205421.1	7	157.02	42.50
OZ205422.1	8	130.92	40.50
OZ205423.1	9	115.72	40.50
OZ205424.1	10	111.65	40.50
OZ205425.1	11	110.79	42
OZ205426.1	12	106.58	43
OZ205427.1	13	96.58	43
OZ205428.1	14	103	41.50
OZ205429.1	15	87.27	45.50
OZ205430.1	16	78.99	41
OZ205431.1	17	78.78	39.50
OZ205432.1	18	68.83	46.50
OZ205433.1	19	65.93	46.50
OZ205434.1	20	55.64	41.50
OZ205419.1	X	150.08	40

For haplotype 1, the estimated QV is 63.7, and for haplotype 2, 66.9. When the two haplotypes are combined, the assembly achieves an estimated QV of 64.9. The *k*-mer completeness is 95.37% for haplotype 1, 94.70% for haplotype 2, and 99.10% for the combined haplotypes (Figure 4).

BUSCO analysis using the cetartiodactyla_odb10 reference set (n = 13 335) identified 97.8% of the expected gene set (single = 95.2%, duplicated = 2.6%) for haplotype 1. The snail plot in Figure 5 summarises the scaffold length distribution and other assembly statistics for haplotype 1. The blob plot in Figure 6 shows the distribution of scaffolds by GC proportion and coverage for haplotype 1.

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 haplotype 1, is **6.8.Q63**.

Genome annotation report

The *Mesoplodon mirus* genome assembly (GCA_964341445.1) was annotated by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 33 806 transcribed mRNAs from 18 422 protein-coding and 3 811 non-coding genes. The average transcript length is 52 734.72 bp, with an average of 1.49 coding transcripts per gene and 10.01 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

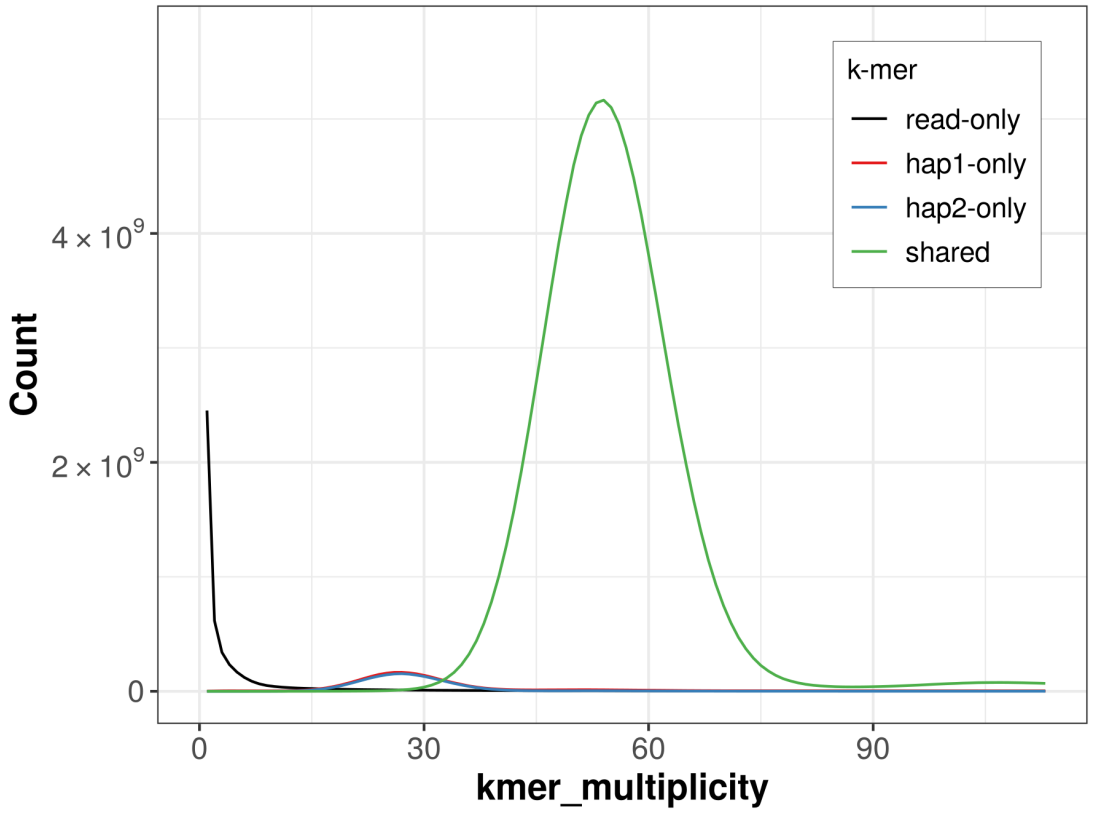

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

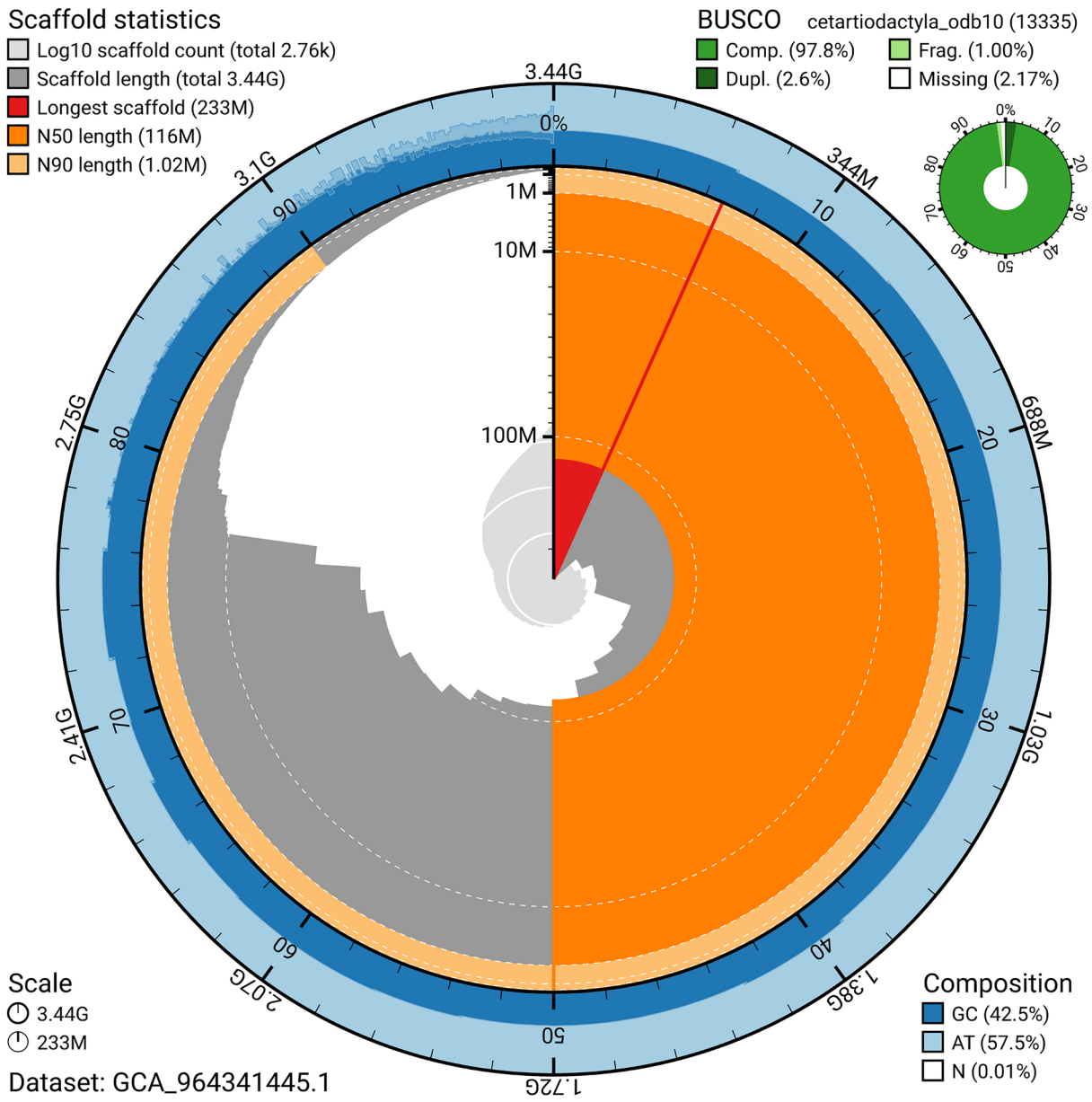

Figure 5. Assembly metrics for mMesMir1.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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.

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)

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,

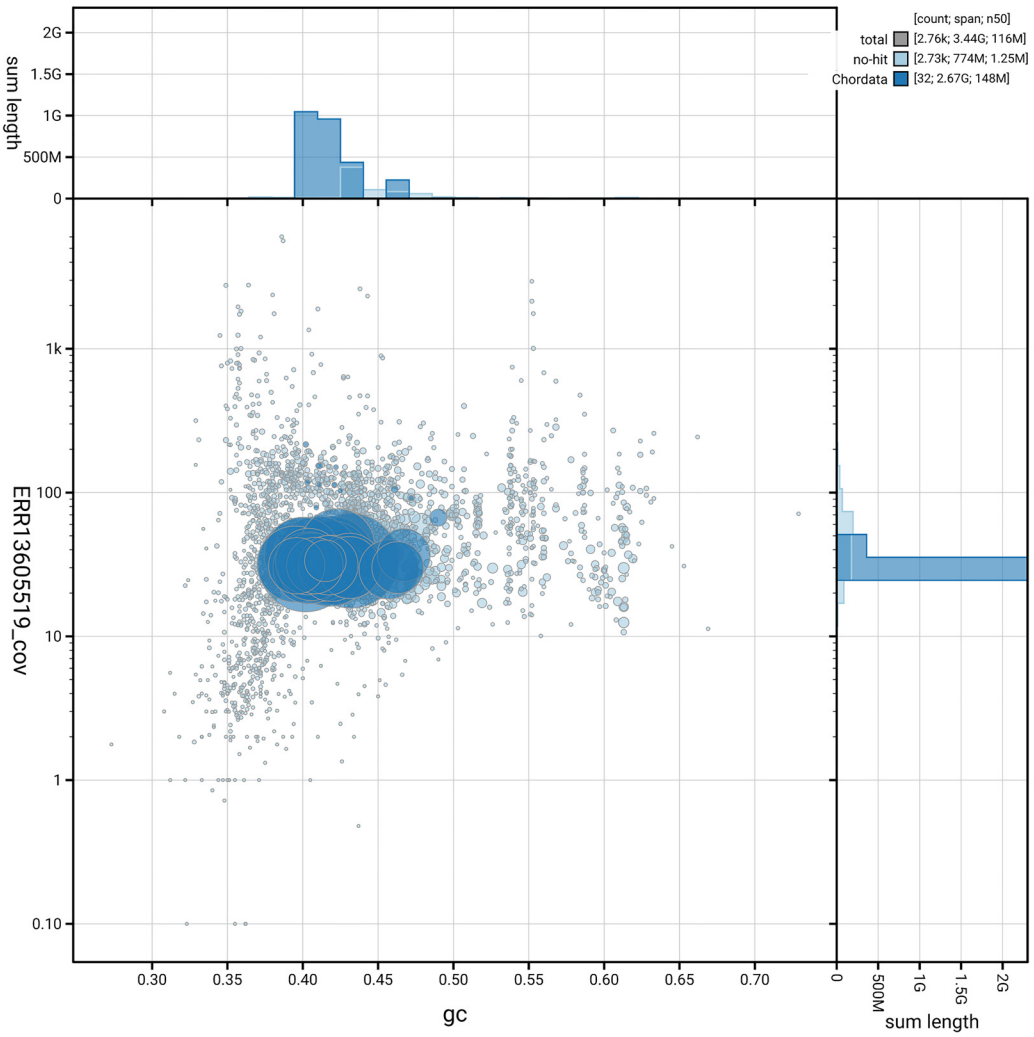

Figure 6. BlobToolKit GC-coverage plot for mMesMir1.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Mesoplodon mirus* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.8.Q63	6.C.Q40
Contig N50 length	3.30 Mb	≥ 1 Mb
Scaffold N50 length	115.72 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 63.7; haplotype 2: 66.9; combined: 64.9	≥ 40
k-mer completeness	Haplotype 1: 95.37%; Haplotype 2: 94.70%; combined: 99.10%	≥ 95%
BUSCO	C:97.8% [S:95.2%; D:2.6%]; F:1.0%; M:1.2%; n:13 335	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	79.62%	≥ 90%

Genome Research Limited (operating as the Wellcome Sanger Institute), and in some circumstances, other Darwin Tree of Life collaborators.

Data availability

European Nucleotide Archive: Mesoplodon mirus (True's beaked whale). Accession number PRJEB79281. The genome sequence is released openly for reuse. The *Mesoplodon mirus* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), Sanger Institute Tree of Life Programme (PRJEB43745), Cetacean Genomes Project (PRJNA1020146) and 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 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+/
BlobToolKit	4.4.6	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.8.3	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	0.0.5	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	v0.8.0	https://github.com/sanger-tol/blobtoolkit

Software	Version	Source
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

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

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

Carroll EL, McGowen MR, McCarthy ML, *et al.*: **Speciation in the deep: genomics and morphology reveal a new species of beaked whale *Mesoplodon eueu*.** *Proc Biol Sci.* 2021; **288**(1961): 20211213.
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

Cheng H, Jarvis ED, Fedrigo O, *et al.*: **Haplotype-resolved assembly of diploid genomes without parental data.** *Nat Biotechnol.* 2022; **40**(9): 1332–1335.
PubMed Abstract | Publisher Full Text | Free Full Text

Committee on Taxonomy, Society for Marine Mammalogy: **List of marine mammal species and subspecies.** 2025.
Reference Source

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

Davison NJ, Morin PA, Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team, *et al.*: **The genome sequence of long-finned pilot whale, *Globicephala melas* (Traill, 1809) [version 1; peer review: 2 approved, 1 approved with reservations].** *Wellcome Open Res.* 2025; **10**: 180.
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

Hernandez-Milian G, Lusher A, O'Brian J, *et al.*: **New information on the diet of True's beaked whale (*Mesoplodon mirus*, Gray 1850), with insights into foraging ecology on mesopelagic prey.** *Mar Mamm Sci.* 2017; **33**(4): 1245–54.
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

Pitman R: ***Mesoplodon* beaked whales: *Mesoplodon spp*.** In: B. Würsig, J. G. M. Thewissen, and K. M. Kovacs (eds), *Encyclopedia of Marine Mammals.* San Diego: Academic Press, 2018; 595–602.
Publisher Full Text

Pitman RL, Brownell RL Jr, Cholewiak D: ***Mesoplodon mirus*.** 2022.
Publisher 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

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 28 January 2026

https://doi.org/10.21956/wellcomeopenres.27999.r144384

✔ **Mariana F Nery**
State University of Campinas, São Paulo, Brazil

This manuscript presents a high-quality chromosomal-level genome assembly for *Mesoplodon mirus* as part of the Darwin Tree of Life Project. The work is technically sound, well-executed, and clearly written. The assembly metrics are excellent, and the methods are described in sufficient detail for reproducibility.

My only doubt is regarding the chromosome assignment percentage. 79,62% of the assembly is assigned to chromosomes, which falls short of the EBP general benchmark. What would comprise the remaining ~20%?

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: 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 19 January 2026

https://doi.org/10.21956/wellcomeopenres.27999.r144382

Jayan Duminda M Senevirathna
Uva Wellassa University, Badulla, Uva Province, Sri Lanka

The manuscript presents a well-executed and clearly written genome assembly for *Mesoplodon mirus*. The authors have produced a high-quality draft assembly supported by strong metrics, including a BUSCO completeness of 97.8%, which is fully acceptable for a vertebrate genome and indicates that the assembly captures the vast majority of expected conserved genes.

The combination of adequate sequencing depth, appropriate assembly tools, and thorough contamination assessment (e.g., blob plot analysis) further supports the reliability of the assembly.

The methods are described in sufficient detail to allow reproducibility, and the results are logically presented and well supported by figures and tables. The manuscript is also strengthened by clear data accessibility statements.

Only minor improvements are suggested, such as adding clarification on annotation parameter settings and briefly discussing the small proportion of missing BUSCOs to provide context regarding potential assembly gaps or repetitive regions. With these minor revisions addressed, the manuscript is suitable for indexing.

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: Aquatic Biotechnology and Genetics

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.