





DATA NOTE

The genome sequence of long-finned pilot whale, *Globicephala melas* (Traill, 1809)

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

Nicholas J. Davison ¹, Phillip A. Morin ²,
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

¹University of Glasgow, Glasgow, Scotland, UK

²Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, La Jolla, California, USA

V1 First published: 08 Apr 2025, 10:180
<https://doi.org/10.12688/wellcomeopenres.23919.1>
Latest published: 08 Apr 2025, 10:180
<https://doi.org/10.12688/wellcomeopenres.23919.1>

Abstract

We present a genome assembly from a male specimen of *Globicephala melas* (long-finned pilot whale; Chordata; Mammalia; Artiodactyla; Delphinidae). The genome sequence has a total length of 2,651.28 megabases. Most of the assembly (89.15%) is scaffolded into 23 chromosomal pseudomolecules, including the X and Y sex chromosomes. The mitochondrial genome has also been assembled, with a length of 16.39 kilobases. Gene annotation of this assembly on Ensembl identified 17,911 protein-coding genes.

Keywords




Globicephala melas, long-finned pilot whale, genome sequence, chromosomal, Artiodactyla



This article is included in the [Tree of Life gateway](#).

Open Peer Review

Approval Status   

	1	2	3
version 1			
08 Apr 2025	view	view	view

1. **Isabella M Reeves** , Flinders University,, Bedford Park, Australia
2. **Takushi Kishida** , Nihon University, Kanagawa, Japan
3. **Janke Axel** , BiK-F/Goethe University/Senckenberg, Frankfurt, Germany

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; Morin PA: Investigation, Project Administration, 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.

Copyright: © 2025 Davison NJ *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: 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 Research 2025, 10:180 <https://doi.org/10.12688/wellcomeopenres.23919.1>

First published: 08 Apr 2025, 10:180 <https://doi.org/10.12688/wellcomeopenres.23919.1>

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; Delphinidae; *Globicephala*; *Globicephala melas* (Traill, 1809) (NCBI:txid9731)

Background

Pilot whales, named for an early theory that schools were piloted by a leader, consist of two recognized species, the short-finned pilot whale (*Globicephala macrorhynchus*), and the long-finned pilot whale (*Globicephala melas*). They have also been referred to as pothead whales (due to the large bulbous forehead or ‘melon’) and blackfish, a term commonly applied to several species with similar characteristics. The short-finned pilot whale is found in tropical, sub-tropical and warm temperate waters globally, potentially consisting of multiple subspecies (Van Cise *et al.*, 2016; Van Cise *et al.*, 2019), though official taxonomic recognition will require additional research. The long-finned pilot whale is distributed in cold temperate waters, with anti-tropically separated subspecies in the North Atlantic (*G. melas melas*) and Southern Hemisphere (*G. melas edwardii*) (Olson, 2018).

Pilot whales are typically nomadic and widely distributed in regions pelagic and coastal or oceanic regions, often associated with areas of high topographic relief and the continental shelf break and slope, with some coastal and island-associated resident populations. Primary diet consists of squid and other cephalopods, and smaller amounts of fish. They are highly social, typically found in schools averaging 20 to 90 individuals, consisting of stable pods of 10–20 individuals with close matrilineal associations (Olson, 2018).

While both long-finned and short-finned pilot whales are considered abundant and listed as Least Concern globally for extinction risk by the IUCN (IUCNredlist.org, consulted 20 September, 2024), they are subject to direct exploitation in some areas, bycatch in fisheries, zoonotic disease, pollution, anthropogenic noise (such as naval sonar and air guns used in oil and gas exploration) and climate change. They are one of the most frequently reported species involved in mass strandings, though the causes of such strandings remains uncertain. Resident populations might be particularly vulnerable to anthropogenic impacts (Van Cise *et al.*, 2017).

The genome of a long-finned pilot whale, *Globicephala melas*, from the North Atlantic subspecies (*G. melas melas*) 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. We present a chromosome-level complete genome sequence for *Globicephala melas*, based on a male specimen from Skye, Scotland, UK.

Genome sequence report

Sequencing data

The genome of a specimen of *Globicephala melas* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 82.54 Gb (gigabases) from 12.10 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 2,739.70 Mb, with a heterozygosity of 0.03% and repeat content of 23.94%. These values provide an initial assessment of genome complexity and the challenges anticipated during assembly. Based on this estimated genome size, the sequencing data provided approximately 29.0x coverage of the genome. Chromosome conformation Hi-C sequencing produced 401.75 Gb from 2,660.60 million reads. Table 1 summarises the specimen and sequencing information.

Assembly statistics

The primary haplotype was assembled, and contigs corresponding to an alternate haplotype were also deposited in INSDC databases. The assembly was improved by manual curation, which corrected 15 misjoins or missing joins. These interventions decreased the scaffold count by 1.0% and increased the scaffold N50 by 9.23%. The final assembly has a total length of 2,651.28 Mb in 992 scaffolds, with 1,084 gaps, and a scaffold N50 of 105.09 Mb (Table 2).

The snail plot in Figure 2 provides a summary of the assembly statistics, indicating the distribution of scaffold lengths and other assembly metrics. Figure 3 shows the distribution of scaffolds by GC proportion and coverage. Figure 4 presents a cumulative assembly plot, with separate curves representing different scaffold subsets assigned to various phyla, illustrating the completeness of the assembly.

Most of the assembly sequence (89.15%) was assigned to 23 chromosomal-level scaffolds, representing 21 autosomes and the



Figure 1. Photograph of the *Globicephala melas* (mGloMel1) specimen used for genome sequencing.

Table 1. Specimen and sequencing data for *Globicephala melas*.

Project information			
Study title	Globicephala melas (long-finned pilot whale)		
Umbrella BioProject	PRJEB64971		
Species	Globicephala melas		
BioSpecimen	SAMEA111380538		
NCBI taxonomy ID	9731		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	mGloMel1	SAMEA111380546	lung
Hi-C sequencing	mGloMel1	SAMEA111380546	lung
RNA sequencing	mGloMel1	SAMEA111380546	lung
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR11837528	2.66e+09	401.75
PacBio Sequel IIe	ERR11843439	2.91e+06	19.21
PacBio Sequel IIe	ERR11843441	3.26e+06	23.65
PacBio Sequel IIe	ERR11843438	3.10e+06	20.47
PacBio Sequel IIe	ERR11843440	2.83e+06	19.2
RNA Illumina NovaSeq 6000	ERR11837529	3.36e+07	5.08

Table 2. Genome assembly data for *Globicephala melas*.

Genome assembly		
Assembly name	mGloMel1.2	
Assembly accession	GCA_963455315.2	
Alternate haplotype accession	GCA_963455345.2	
Assembly level for primary assembly	chromosome	
Span (Mb)	2,651.28	
Number of contigs	2,076	
Number of scaffolds	992	
Longest scaffold (Mb)	188.11	
Assembly metric	Measure	Benchmark
Contig N50 length	3.29 Mb	≥ 1 Mb
Scaffold N50 length	105.09 Mb	= chromosome N50
Consensus quality (QV)	Primary: 61.6; alternate: 60.7; combined: 61.2	≥ 40
k-mer completeness	Primary: 97.07%; alternate: 77.15%; combined: 99.44%.	≥ 95%
BUSCO*	C:95.5%[S:93.3%,D:2.2%], F:1.0%,M:3.5%,n:13,335	S > 90%; D < 5%
Percentage of assembly mapped to chromosomes	89.15%	≥ 90%
Sex chromosomes	X and Y	localised homologous pairs
Organelles	Mitochondrial genome: 16.39 kb	complete single alleles

* BUSCO scores based on the cetartiodactyla_odb10 BUSCO set using version 5.5.0. C = complete [S = single copy, D = duplicated], F = fragmented, M = missing, n = number of orthologues in comparison.

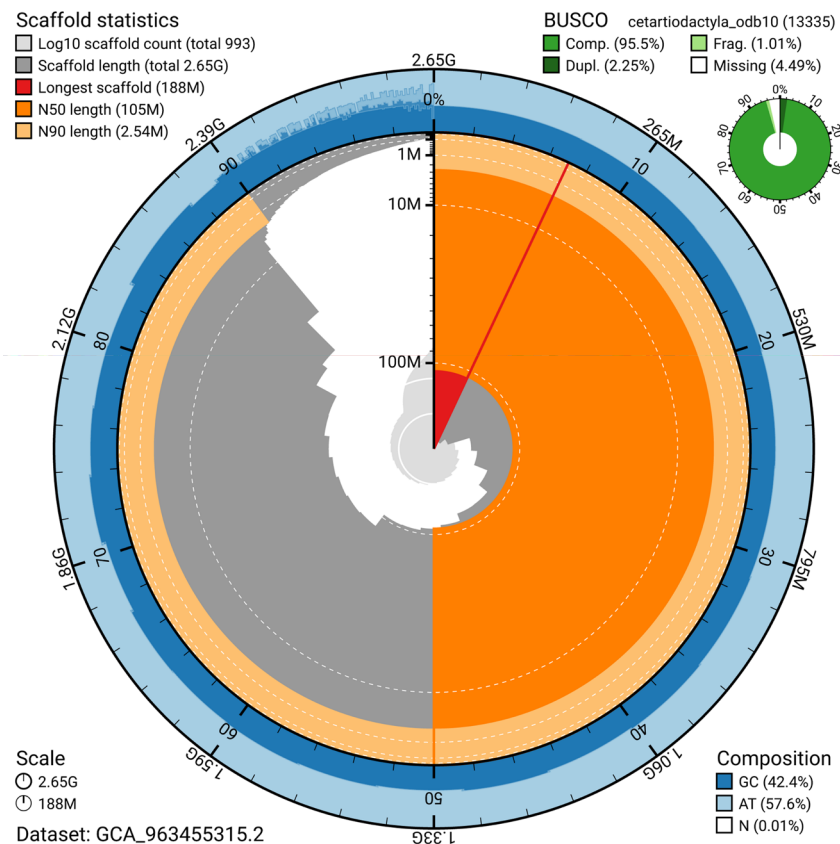


Figure 2. Genome assembly of *Globicephala melas*, mGloMel1.2: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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 cetartiodactyla_odb10 set is presented at the top right. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963455315.2/dataset/GCA_963455315.2/snail.

X and Y sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, it was noted that Chromosomes X and Y were assigned by read coverage statistics and synteny to the genome of *Delphinus delphis* (GCA_949987515.1).

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.

Assembly quality metrics

The estimated Quality Value (QV) and *k*-mer completeness metrics, along with BUSCO completeness scores, were calculated for each haplotype and the combined assembly. The QV reflects the base-level accuracy of the assembly, while *k*-mer completeness indicates the proportion of expected *k*-mers identified in the assembly. BUSCO scores provide a measure of completeness based on benchmarking universal single-copy orthologues.

The combined primary and alternate assemblies achieve an estimated QV of 61.2. The *k*-mer recovery for the primary haplotype is 97.07%, and for the alternate haplotype 77.15%; the combined primary and alternate assemblies have a *k*-mer recovery of 99.44%. BUSCO v.5.5.0 analysis using the cetartiodactyla_odb10 reference set ($n = 13,335$) identified 95.5% of the expected gene set (single = 93.3%, duplicated = 2.2%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project (EBP) Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of **6.8.Q61**.

Genome annotation report

The *Globicephala melas* genome assembly (GCA_963455315.1) was annotated externally by Ensembl at the European Bioinformatics Institute (EBI). This annotation includes 39,375 transcribed mRNAs from 17,911 protein-coding and 5,689 non-coding genes. The average transcript length is 60,316.82.

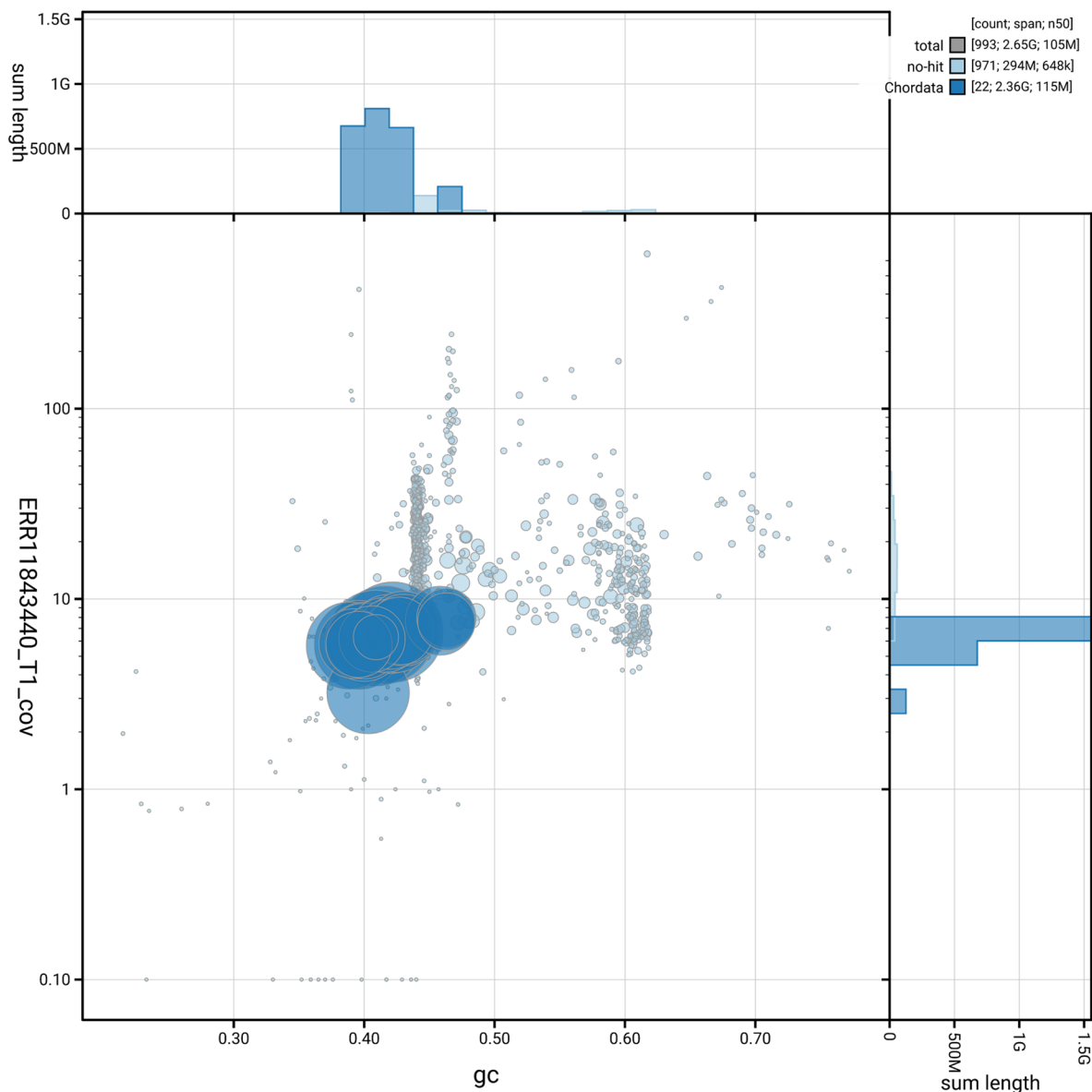


Figure 3. Genome assembly of *Globicephala melas*, mGloMe1.2: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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 at https://blobtoolkit.genomehubs.org/view/GCA_963455315.2/dataset/GCA_963455315.2/blob.

There are 1.64 coding transcripts per gene and 9.84 exons per transcript. For further information about the annotation, please refer to <https://beta.ensembl.org/species/cf40824d-11c6-4bb7-aff2-18ecce2bac7a>.

Methods

Sample acquisition

An adult male *Globicephala melas* (specimen ID SAN00002603, ToLID mGloMe1) was collected from Coruisk, Loch Na Cuilce, Skye, Highland, Scotland (latitude 57.1989, longitude -6.165) on 2022-05-03. The specimen was collected and

identified by Nick Davison (Scottish Marine Animal Stranding Scheme University of Glasgow). A sample of lung was collected at necropsy and preserved by freezing at -80 °C.

Metadata collection for samples adhered to the Darwin Tree of Life project standards described by [Lawniczak et al. \(2022\)](#).

Nucleic acid extraction

The workflow for high molecular weight (HMW) DNA extraction at the Wellcome Sanger Institute (WSI) Tree of Life Core Laboratory includes a sequence of procedures: sample

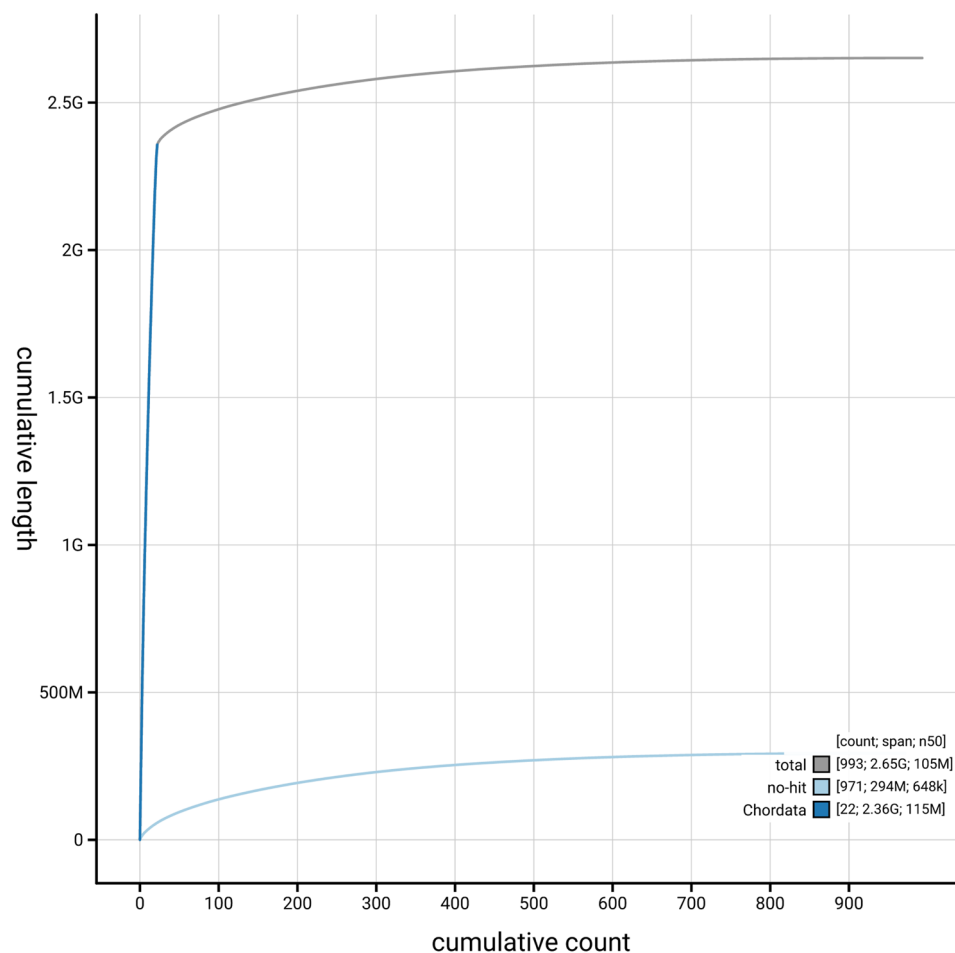


Figure 4. Genome assembly of *Globicephala melas*, mGloMel1.2: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963455315.2/dataset/GCA_963455315.2/cumulative.

preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023). The mGloMel1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the lung was cryogenically disrupted using the Covaris cryoPREP® Automated Dry Pulverizer (Narváez-Gómez *et al.*, 2023). HMW DNA was extracted using the Automated MagAttract v1 protocol (Sheerin *et al.*, 2023). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Todorovic *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Strickland *et al.*, 2023). The concentration of the sheared and purified DNA was assessed using a Nanodrop spectrophotometer and a Qubit Fluorometer using the Qubit dsDNA High Sensitivity Assay kit. The fragment size distribution was evaluated by running the sample on the FemtoPulse system.

RNA was extracted from lung tissue of mGloMel1 in the Tree of Life Laboratory at the WSI using the RNA Extraction: Automated MagMax™ mirVana protocol (do Amaral *et al.*, 2023). 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.

Hi-C sample preparation and crosslinking

Tissue from the lung of the mGloMel1 sample was processed for Hi-C sequencing at the WSI Scientific Operations core, using the Arima-HiC v2 kit. In brief, 20–50 mg of frozen tissue (stored at –80 °C) was fixed, and the DNA crosslinked using a TC buffer with 22% formaldehyde concentration. After crosslinking, the tissue was homogenised using the Diagnocine Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked

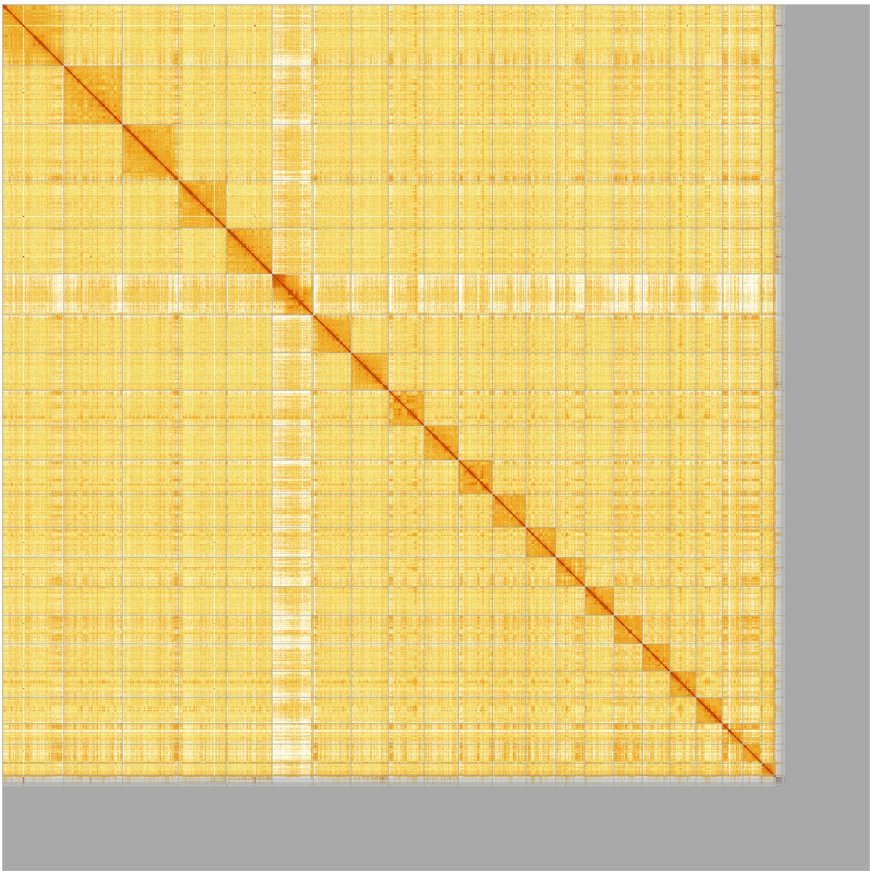


Figure 5. Genome assembly of *Globicephala melas*: Hi-C contact map of the mGloMel1.2 assembly, produced in PretextView. Chromosomes are shown in order of size from left to right and top to bottom.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Globicephala melas*, mGloMel1.

INSDC accession	Name	Length (Mb)	GC%
OY734039.1	1	188.11	42
OY734040.2	2	178.82	41.5
OY734041.1	3	171.32	41
OY734042.1	4	146.48	39.5
OY734043.1	5	140.22	39
OY734045.1	6	115.44	42
OY734046.1	7	115.18	40
OY734047.1	8	108.12	42.5
OY734048.1	9	105.09	40
OY734049.1	10	103.04	41.5
OY734050.2	11	102.7	43

INSDC accession	Name	Length (Mb)	GC%
OY734051.1	12	90.74	42
OY734052.1	13	89.85	43
OY734053.1	14	88.67	39
OY734054.1	15	86.73	46
OY734055.1	16	83.95	42.5
OY734056.1	17	79.95	40.5
OY734057.1	18	79.7	39.5
OY734058.1	19	62.92	46.5
OY734059.1	20	58.78	46.5
OY734060.1	21	35.6	41
OY734044.1	X	125.8	40.5
OY734061.1	Y	6.52	42
OY734062.1	MT	0.02	39

DNA was digested using a restriction enzyme master mix. The 5'-overhangs were filled in and labelled with biotinylated nucleotides and proximally ligated. An overnight incubation was carried out for enzymes to digest remaining proteins and for crosslinks to reverse. A clean up was performed with SPRIselect beads prior to library preparation. Additionally, the biotinylation percentage was estimated using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) and Qubit HS Assay Kit and Arima-HiC v2 QC beads.

Library preparation and sequencing

Library preparation and sequencing were performed at the WSI Scientific Operations core.

PacBio HiFi

At a minimum, samples were required to have an average fragment size exceeding 8 kb and a total mass over 400 ng to proceed to the low input SMRTbell Prep Kit 3.0 protocol (Pacific Biosciences, California, USA), depending on genome size and sequencing depth required. Libraries were prepared using the SMRTbell Prep Kit 3.0 (Pacific Biosciences, California, USA) as per the manufacturer's instructions. The kit includes the reagents required for end repair/A-tailing, adapter ligation, post-ligation SMRTbell bead cleanup, and nuclease treatment. Following the manufacturer's instructions, size selection and clean up was carried out using diluted AMPure PB beads (Pacific Biosciences, California, USA). DNA concentration was quantified using the Qubit Fluorometer v4.0 (Thermo Fisher Scientific) with Qubit 1X dsDNA HS assay kit and the final library fragment size analysis was carried out using the Agilent Femto Pulse Automated Pulsed Field CE Instrument (Agilent Technologies) and gDNA 55kb BAC analysis kit.

Samples were 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, as well as perform primary and secondary analysis of the data upon completion.

Hi-C

For Hi-C library preparation, DNA was fragmented using the Covaris E220 sonicator (Covaris) and size selected using SPRIselect beads to 400 to 600 bp. The DNA was then enriched using the Arima-HiC v2 kit Enrichment beads. Using the NEBNext Ultra II DNA Library Prep Kit (New England Biolabs) for end repair, A-tailing, and adapter ligation. This uses a custom protocol which resembles the standard NEBNext Ultra II DNA Library Prep protocol but where library preparation occurs while DNA is bound to the Enrichment beads. For library amplification, 10 to 16 PCR cycles were required, determined by the sample biotinylation percentage. The Hi-C sequencing was performed using paired-end sequencing with a read length of 150 bp on an Illumina NovaSeq 6000 instrument.

RNA

Poly(A) RNA-Seq libraries were constructed using the NEB Ultra II RNA Library Prep kit, following the manufacturer's instructions. RNA sequencing was performed on the Illumina NovaSeq 6000 instrument.

Genome assembly, curation and evaluation

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 first 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 using 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. Flat files and maps used in curation were generated via the TreeVal pipeline (Pinton *et al.*, 2023). Manual curation was conducted primarily in PretextView (Harry, 2022) and HiGlass (Kerpedjiev *et al.*, 2018), with additional insights provided by JBrowse2 (Diesh *et al.*, 2023). Scaffolds were visually inspected and corrected as described by Howe *et al.* (2021). Any identified contamination, missed joins, and mis-joins were amended, and duplicate sequences were tagged and removed. The curation process is documented at <https://gitlab.com/wtsi-grit/rapid-curation>.

Assembly quality assessment

The Merqury.FK tool (Rhie *et al.*, 2020), run in a Singularity container (Kurtzer *et al.*, 2017), was used 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 blobtoolkit pipeline is a Nextflow (Di Tommaso *et al.*, 2017) port of the previous Snakemake Blobtoolkit pipeline (Challis *et al.*, 2020). It aligns the PacBio reads in SAMtools

and minimap2 (Li, 2018) and generates coverage tracks for regions of fixed size. In parallel, it queries the GoAT database (Challis *et al.*, 2023) to identify all matching BUSCO lineages to run BUSCO (Manni *et al.*, 2021). For the three domain-level BUSCO lineages, the pipeline aligns the BUSCO genes to the UniProt Reference Proteomes database (Bateman *et al.*, 2023) with DIAMOND blastp (Buchfink *et al.*, 2021). The genome is also divided into chunks according to the density of the BUSCO genes from the closest taxonomic lineage, and each chunk is aligned to the UniProt Reference Proteomes database using DIAMOND blastx. Genome sequences without a hit are chunked using seqtk and aligned to the NT database with blastn (Altschul *et al.*, 1990). The blobtools suite combines all these 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), relying on the Conda package manager, the Bioconda initiative (Grüning *et al.*, 2018), the Biocontainers infrastructure (da Veiga Leprevost *et al.*, 2017), as well as the Docker (Merkel, 2014) and Singularity (Kurtzer *et al.*, 2017) containerisation solutions.

Table 4 contains a list of relevant software tool versions and sources.

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

Table 4. Software tools: versions and sources.

Software tool	Version	Source
BLAST	2.14.0	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/
BlobToolKit	4.3.9	https://github.com/blobtoolkit/blobtoolkit
BUSCO	5.5.0	https://gitlab.com/ezlab/busco
bwa-mem2	2.2.1	https://github.com/bwa-mem2/bwa-mem2
DIAMOND	2.1.8	https://github.com/bbuchfink/diamond
fasta_windows	0.2.4	https://github.com/tolkite/fasta_windows
FastK	666652151335353eef2fcd58880bcef5bc2928e1	https://github.com/thegenemyers/FASTK
Gfastats	1.3.6	https://github.com/vgl-hub/gfastats
GoAT CLI	0.2.5	https://github.com/genomehubs/goat-cli
Hifiasm	0.19.5-r587	https://github.com/chhylp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84aa44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquyFK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQUERY.FK
Minimap2	2.24-r1122	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	23.10.0	https://github.com/nextflow-io/nextflow
PretextView	0.2.5	https://github.com/sanger-tol/PretextView
purge_dups	1.2.5	https://github.com/dfguan/purge_dups
samtools	1.19.2	https://github.com/samtools/samtools
sanger-tol/ascc	-	https://github.com/sanger-tol/ascc
sanger-tol/blobtoolkit	0.6.0	https://github.com/sanger-tol/blobtoolkit
Seqtk	1.3	https://github.com/lh3/seqtk
Singularity	3.9.0	https://github.com/sylabs/singularity
TreeVal	1.2.0	https://github.com/sanger-tol/treeval
YaHS	1.2a.2	https://github.com/c-zhou/yahs

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 [here](#). 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)

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: *Globicephala melas* (long-finned pilot whale). Accession number PRJEB64971; <https://identifiers.org/ena.embl/PRJEB64971>. The genome sequence is released openly for reuse. The *Globicephala melas* genome sequencing initiative is part of the Darwin Tree of Life (DTOL) project. 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](#).

Author information

Members of the Wellcome Sanger Institute Tree of Life Management, Samples and Laboratory team are listed here: <https://doi.org/10.5281/zenodo.12162482>.

Members of Wellcome Sanger Institute Scientific Operations: Sequencing Operations are listed here: <https://doi.org/10.5281/zenodo.12165051>.

Members of the Wellcome Sanger Institute Tree of Life Core Informatics team are listed here: <https://doi.org/10.5281/zenodo.12160324>.

Members of the Tree of Life Core Informatics collective are listed here: <https://doi.org/10.5281/zenodo.12205391>.

Members of the Darwin Tree of Life Consortium are listed here: <https://doi.org/10.5281/zenodo.4783558>.

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](#)
- Challis R, Kumar S, Sotero-Caio C, *et al.*: **Genomes on a Tree (GoAT): a versatile, scalable search engine for genomic and sequencing project metadata across the eukaryotic Tree of Life [version 1; peer review: 2 approved]**. *Wellcome Open Res.* 2023; **8**: 24. [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](#)
- da Veiga Leprevost F, Gruning BA, Alves Aflitos S, *et al.*: **BioContainers: an open-source and community-driven framework for software standardization**. *Bioinformatics.* 2017; **33**(16): 2580–2582. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Denton A, Yatsenko H, Jay J, *et al.*: **Sanger Tree of Life wet laboratory protocol collection V.1**. *protocols.io.* 2023. [Publisher Full Text](#)
- Di Tommaso P, Chatzou M, Floden EW, *et al.*: **Nextflow enables reproducible computational workflows**. *Nat Biotechnol.* 2017; **35**(4): 316–319. [PubMed Abstract](#) | [Publisher Full Text](#)
- Diesh C, Stevens GJ, Xie P, *et al.*: **JBrowse 2: a modular genome browser with views of synteny and structural variation**. *Genome Biol.* 2023; **24**(1): 74. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- do Amaral RJV, Bates A, Denton A, *et al.*: **Sanger Tree of Life RNA extraction: automated MagMax™ mirVana**. *protocols.io.* 2023. [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](#)
- Gruning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive**

software distribution for the life sciences. *Nat Methods*. 2018; **15**(7): 475–476.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free 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](#)

Harry E: PretextView (Paired REad TEXTure Viewer): a desktop application for viewing pretext contact maps. 2022.
[Reference Source](#)

Howe K, Chow W, Collins J, *et al.*: Significantly improving the quality of genome assemblies through curation. *GigaScience*. 2021; **10**(1): gaa153.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Jay J, Yatsenko H, Narváez-Gómez JP, *et al.*: Sanger Tree of Life sample preparation: triage and dissection. *protocols.io*. 2023.
[Publisher 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, [Accessed 2 April 2024].
[Reference Source](#)

Narváez-Gómez JP, Mbye H, Oatley G, *et al.*: Sanger Tree of Life sample homogenisation: Covaris cryoPREP® automated dry pulverizer V.1. *protocols.io*. 2023.

[Publisher Full Text](#)

Olson PA: Pilot whales: *Globicephala melas* and *G. macrorhynchus*. In: Würsig, B., Thewissen, J. G. M., and Kovacs, K. M. (eds.) *Encyclopedia of marine mammals*. 3rd ed. San Diego: Academic Press, 2018; 701–705.
[Publisher Full Text](#)

Pointon DL, Eagles W, Sims Y, *et al.*: sanger-tol/treeval v1.0.0 – Ancient Atlantis. 2023.
[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](#)

Sheerin E, Sampaio F, Oatley G, *et al.*: Sanger Tree of Life HMW DNA extraction: automated MagAttract v.1. *protocols.io*. 2023.
[Publisher Full Text](#)

Strickland M, Cornwell C, Howard C: Sanger Tree of Life fragmented DNA clean up: manual SPRI. *protocols.io*. 2023.
[Publisher Full Text](#)

Todorovic M, Sampaio F, Howard C: Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor®3 for PacBio HiFi. *protocols.io*. 2023.
[Publisher Full Text](#)

Uliano-Silva M, Ferreira JGRN, Krashenninnikova 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](#)

Van Cise AM, Baird RW, Baker CS, *et al.*: Oceanographic barriers, divergence, and admixture: phylogeography and taxonomy of two putative subspecies of short-finned pilot whale. *Mol Ecol*. 2019; **28**(11): 2886–2902.

[PubMed Abstract](#) | [Publisher Full Text](#)

Van Cise AM, Martien KK, Mahaffy SD, *et al.*: Familial social structure and socially driven genetic differentiation in Hawaiian short-finned pilot whales. *Mol Ecol*. 2017; **26**(23): 6730–6741.

[PubMed Abstract](#) | [Publisher Full Text](#)

Van Cise A, Morin PA, Baird RW, *et al.*: Redrawing the map: mtDNA provides new insight into the distribution and diversity of short-finned pilot whales in the Pacific Ocean. *Mar Mamm Sci*. 2016; **32**(4): 1177–1199.

[Publisher 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 19 April 2025

<https://doi.org/10.21956/wellcomeopenres.26388.r122148>

© 2025 Axel J. 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.



Janke Axel 

Ecological Genomics, BiK-F/Goethe University/Senckenberg, Frankfurt, Germany

Peer Review Comments - MAJOR REVISION

This manuscript presents a genome report which, while lacking a specific biological research question, provides a technically sound account of genome sequencing, assembly, and annotation. The work is professionally conducted, and the methodology is mostly described with sufficient clarity.

However, several important aspects are either missing or require further elaboration to ensure the study's utility for the broader community, particularly for those working on similar non-model or degraded samples.

1. DNA and RNA Quality Assessment:

There is no information provided on the quality or integrity of the extracted DNA. Given that the sample appear to originate from a naturally deceased animal, with an unknown post-mortem interval, this is of critical importance. Please include the Femto-Pulse report or equivalent data showing DNA fragment size distribution. This information is vital to assess the potential limitations of the assembly and to inform future efforts involving similarly compromised material. The same applies to the RNA quality: please provide fragment length distribution and RIN-like metrics for evaluation of transcript integrity. These details are particularly valuable when fresh or high-quality material is not available.

2. Blob Plot Interpretation:

The Blob plot reveals a number of non-target sequences, likely corresponding to bacterial or fungal contamination. While such findings are not unexpected given the possible post-mortem degradation, the manuscript should briefly elaborate on the origin, taxonomy, and abundance of these non-mammalian sequences. It would be helpful to state whether these were filtered prior to assembly or if any were incorporated into scaffolds. This information may help guide decontamination strategies in similar projects.

3. Sample Context and Decay Estimate:

To support the interpretation of the data, please provide an estimate—however approximate—of how long the individual had been deceased prior to sampling. Environmental context (e.g., temperature, exposure, humidity) should also be described to the best of the authors' knowledge. This would help frame the state of preservation and assist others in evaluating feasibility for

related work using non-ideal specimens.

4. Chromosome-Specific Concerns (Chromosome 6):

The heatmap corresponding to the sixth(?) chromosome suggests structural anomalies or misassemblies. Please describe whether any steps were taken investigate or correct these issues (e.g., manual curation, re-scaffolding, use of alternative data types). If no specific action was taken, please clarify why. Similarly, assemblies not in the heatmap should be briefly addressed: were they omitted due to quality, completeness, or other reasons? Even a short rationale would be appreciated.

5. Background Section:

The current background is overly detailed relative to the scope of the manuscript. As no biological analyses or questions are pursued, it would be preferable to reduce this section to the essential taxonomic and minimal biological context required to understand the importance of the project.

Conclusion:

While the manuscript lacks any biological hypothesis (read: it is boring), it still holds potential value for the community, especially as a reference for genome sequencing from compromised material. By addressing the issues above, the manuscript will be significantly improved and maybe more useful except for another genome . Still, I am looking forward to seeing it indexed.

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

Partly

Are the protocols appropriate and is the work technically sound?

Partly

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

Partly

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

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Translational Biodiversity 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, however I have significant reservations, as outlined above.

Reviewer Report 19 April 2025

<https://doi.org/10.21956/wellcomeopenres.26388.r122154>

© 2025 Kishida T. 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.



Takushi Kishida 

Nihon University, Kanagawa, Japan

This study provides a high-quality chromosomal-level long-finned pilot whale haploid genome assembly. I have no complaints about the quality of this genome data.

I hope, if possible, the authors deposit the specimen used for genome sequencing in an appropriate museum so that a voucher specimen associated with this genome data is available.

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 genomics, population 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.

Reviewer Report 18 April 2025

<https://doi.org/10.21956/wellcomeopenres.26388.r122151>

© 2025 Reeves I. 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.



Isabella M Reeves 

Flinders University,, Bedford Park, Australia

This manuscript is an exciting contribution to our field, providing a high-quality genome assembly of the long-finned pilot whale for future research to use. The introduction, which covers the natural history of pilot whales, was particularly engaging.

One suggestion I would like to offer is the inclusion of more detailed information regarding the software and tools used during the genome assembly process. Specifically, it would be beneficial to provide the versions of the software employed, as well as additional details on the parameters/thresholds used throughout the assembly.

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?

Partly

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

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Evolution, genomics, marine mammals, ecology

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.
