Check for updates

DATA NOTE

The genome sequence of the fin whale, *Balaenoptera physalus* (Linnaeus, 1758) (Artiodactyla: Balaenopteridae)

[version 1; peer review: 2 approved]

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

[1]University of Glasgow, Glasgow, Scotland, UK
[2]Southwest Fisheries Science Center, NOAA, La Jolla, California, USA

v1 **First published:** 05 Jan 2026, **11**:5
https://doi.org/10.12688/wellcomeopenres.25412.1
Latest published: 05 Jan 2026, **11**:5
https://doi.org/10.12688/wellcomeopenres.25412.1

Abstract

We present a genome assembly from an individual male *Balaenoptera physalus* (fin whale; Chordata; Mammalia; Artiodactyla; Balaenopteridae). The assembly contains two haplotypes with total lengths of 3 442.54 megabases and 2 850.21 megabases. Most of haplotype 1 (79.11%) is scaffolded into 23 chromosomal pseudomolecules, including the X and Y sex chromosomes. Haplotype 2 was assembled to scaffold level. The mitochondrial genome has also been assembled, with a length of 16.4 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

Balaenoptera physalus; fin whale; genome sequence; chromosomal; Artiodactyla

This article is included in the Tree of Life gateway.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 1 05 Jan 2026	✓ view	✓ view

1. **Mariana F Nery**, State University of Campinas, São Paulo, Brazil
2. **L Filipe C Castro**, University of Porto, Porto, Portugal

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

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

Author roles: Davison NJ: Investigation, Resources, Writing – Review & Editing; **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.

Copyright: © 2026 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 the fin whale, *Balaenoptera physalus* (Linnaeus, 1758) (Artiodactyla: Balaenopteridae) [version 1; peer review: 2 approved]** Wellcome Open Research 2026, **11**:5 https://doi.org/10.12688/wellcomeopenres.25412.1

First published: 05 Jan 2026, **11**:5 https://doi.org/10.12688/wellcomeopenres.25412.1

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; Mysticeti; Balaenopteridae; *Balaenoptera*; *Balaenoptera physalus* (Linnaeus, 1758) (NCBI:txid9770)

Background

The fin whale (*Balaenoptera physalus)* is the second largest animal on earth (up to 23–26 m in length), only slightly smaller than its close relative, the blue whale (*Balaenoptera musculus*) (up to 33 m in length). Fin whales are antitropically distributed in temperate to subpolar waters, following seasonal migrations that isolate populations in each hemisphere, and are subdivided into three oceanic subspecies in the North Pacific (*B. p. velifera*), North Atlantic (*B. p. physalus*), and the Southern Ocean (*B. p. quoyi*) (Archer *et al.*, 2019).

Fin whales are baleen whales that lunge- and filter-feed on a variety of small schooling fish, squid, and krill, varying by season and locality. They swim fast (5 to 8 knots, up to 15 knots) and cover large distances in search of food and during seasonal migrations, typically traveling singly or in small groups. Maturation occurs at around 25 years, and individuals have been known to live for 80 to 90 years (Aguilar & García-Vernet, 2018).

More fin whales were killed than any other species during industrial whaling, with approximately 874 000 fin whales killed in the 20th century (Rocha *et al.*, 2014), reducing global abundance by ~70% (Edwards *et al.*, 2015), and some populations by ~99% (Nigenda-Morales *et al.*, 2023). Fin whales are listed as Vulnerable globally, and Endangered in the Mediterranean Sea (Cooke, 2018). Although commercial whaling effectively stopped in the 1980s due to the global moratorium, some aboriginal (Greenland) and commercial whaling (Japan, Iceland) of fin whales has continued. Other threats include ship strikes, entanglement in commercial fishing gear, pollution, and climate change.

Two scaffold-level genome assemblies have previously been generated for *Balaenoptera physalus* (GCA_023338255.1, GCA_008795845.1; submitted by Senckenberg Bik-F and the Korea Ocean Research & Development Institute) (data obtained via NCBI datasets, O'Leary *et al.*, 2024). We present a chromosome-level genome sequence for the species, produced using the Tree of Life pipeline from a specimen collected from Orkney, Scotland, UK (Figure 1).

Methods

Sample acquisition

The specimen used for genome sequencing was an adult male *Balaenoptera physalus* (specimen ID SAN00003288, ToLID mBalPhy2; Figure 1), collected from Echnaloch Bay, Burray, Orkney, Scotland, United Kingdom (latitude 58.8568, longitude –2.9242) on 2021-07-21. The specimen was collected and identified by Nick Davison (Scottish Marine Animal Stranding Scheme, University of Glasgow). Sample metadata were collected in line with the Darwin Tree of Life project standards described by Lawniczak *et al.* (2022).

Figure 1. Members of the Scottish Marine Animal Stranding Scheme with the *Balaenoptera physalus* (mBalPhy2) carcass from which samples were taken for genome sequencing (photo credit Nick Davison).

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 mBalPhy2 sample was weighed and triaged to determine the appropriate extraction protocol. Tissue from the muscle 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.

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 from the mBalPhy2 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.

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 33 breaks and 171 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 *Balaenoptera physalus* specimen generated 82.14 Gb (gigabases) from 8.00 million reads, which were used to assemble the genome. GenomeScope2.0 analysis estimated the haploid genome size at 3 160.70 Mb, with a heterozygosity of 0.26% and repeat content of 36.13% (Figure 2). These estimates guided expectations for the assembly. Based on the estimated genome size, the sequencing data provided approximately 25× coverage. Hi-C sequencing produced 406.50 Gb from 2 692.03 million reads, which were used to scaffold the assembly. 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 has a total length of 3 442.54 Mb in 2 611 scaffolds, with 1 516 gaps, and a scaffold N50 of 110.87 Mb (Table 2).

Most of the assembly sequence (79.11%) was assigned to 23 chromosomal-level scaffolds, representing 21 autosomes and the X and Y sex chromosomes. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 3; Table 3). Scaffolds in the following regions have highly repetitive sequence, and have uncertain order and orientation: 20 to 22.5 Mbp on Chromosome 4, 22.4 to 32.3 and 108.1 to 119.4 Mbp on Chromosome 6, 26 to 42.3 Mbp on Chromosome 7, 73 to 88.4 Mbp on Chromosome 10, 88.1 to 97.2 Mbp on Chromosome 13, 34.5 to 39.1 Mbp on Chromosome 17, 23.7 to 29.6 Mbp on Chromosome 19, 1.2 to 7.7 Mbp on Chromosome 21 and 18 to 24.3 Mbp on Chromosome Y. The sex chromosomes were identified by the depth of PacBio read coverage.

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.

For haplotype 1, the estimated QV is 65.6, and for haplotype 2, 65.5. When the two haplotypes are combined, the assembly achieves an estimated QV of 65.6. The *k*-mer completeness is 94.59% for haplotype 1, 89.62% for haplotype 2, and 99.16% for the combined haplotypes (Figure 4).

BUSCO analysis using the cetartiodactyla_odb10 reference set (n = 13 335) identified 95.6% of the expected gene set (single = 92.7%, duplicated = 2.9%) 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

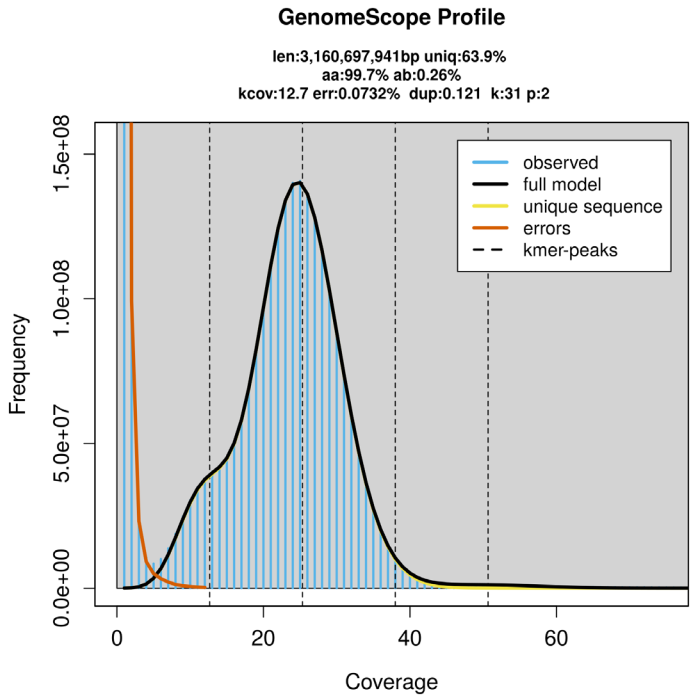

Figure 2. Frequency distribution of *k*-mers generated using GenomeScope2. The plot shows observed and modelled *k*-mer spectra, providing estimates of genome size, heterozygosity, and repeat content based on unassembled sequencing reads.

Table 1. Specimen and sequencing data for BioProject PRJEB76339.

Platform	PacBio HiFi	Hi-C
ToLID	mBalPhy2	mBalPhy2
Specimen ID	SAN00003288	SAN00003288
BioSample (source individual)	SAMEA114493136	SAMEA114493136
BioSample (tissue)	SAMEA114493144	SAMEA114493144
Tissue	muscle	muscle
Instrument	Revio	Illumina NovaSeq X
Run accessions	ERR13245265	ERR13248928
Read count total	8.00 million	2 692.03 million
Base count total	82.14 Gb	406.50 Gb

Table 2. Genome assembly statistics.

Assembly name	mBalPhy2.hap1.1	mBalPhy2.hap2.1
Assembly accession	GCA_965194825.1	GCA_965194765.1
Assembly level	chromosome	scaffold
Span (Mb)	3 442.54	2 850.21
Number of chromosomes	23	scaffold-level
Number of contigs	4 127	3 706
Contig N50	2.08 Mb	2.31 Mb
Number of scaffolds	2 611	2 448
Scaffold N50	110.87 Mb	106.92 Mb
Longest scaffold length (Mb)	197.29	-
Sex chromosomes	X and Y	-
Organelles	Mitochondrion: 16.4 kb	-

Report on Assembly Standards September 2024. The EBP metric, calculated for the haplotype 1, is **6.8.Q65**.

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)

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,

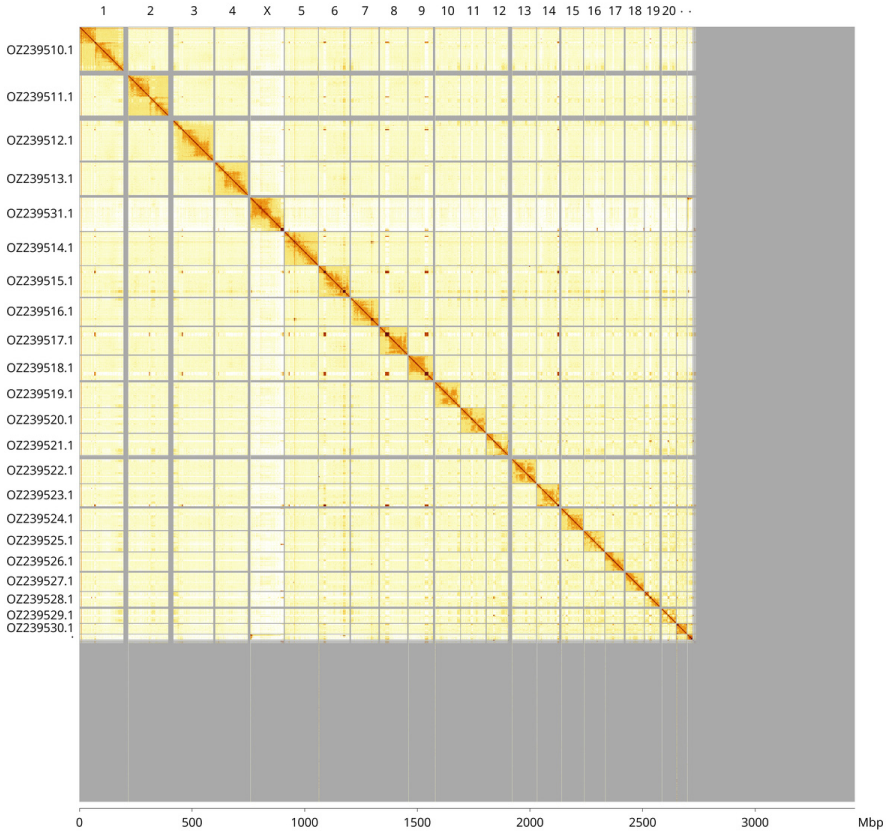

Figure 3. Hi-C contact map of the *Balaenoptera physalus* genome assembly. Assembled chromosomes are shown in order of size and labelled along the axes, with a megabase scale shown below. The plot was generated using PretextSnapshot.

Table 3. Chromosomal pseudomolecules in the haplotype 1 genome assembly of *Balaenoptera physalus* mBalPhy2.

INSDC accession	Molecule	Length (Mb)	GC%
OZ239510.1	1	215.55	41.50
OZ239511.1	2	200.59	41
OZ239512.1	3	185.94	40.50
OZ239513.1	4	155.67	39
OZ239514.1	5	151.80	38.50
OZ239515.1	6	142.26	40.50
OZ239516.1	7	128.10	41.50
OZ239517.1	8	128.02	40
OZ239518.1	9	117.78	39.50
OZ239519.1	10	114.40	42
OZ239520.1	11	114.32	43
OZ239521.1	12	113.45	43
OZ239522.1	13	109.76	41.50
OZ239523.1	14	108.97	43
OZ239524.1	15	100.29	41.50
OZ239525.1	16	93.56	45.50
OZ239526.1	17	90.56	40.50
OZ239527.1	18	85.40	39
OZ239528.1	19	75.79	45
OZ239529.1	20	66.03	46
OZ239530.1	21	47.46	41.50
OZ239531.1	X	152.38	40.50
OZ239532.1	Y	25.40	41.50

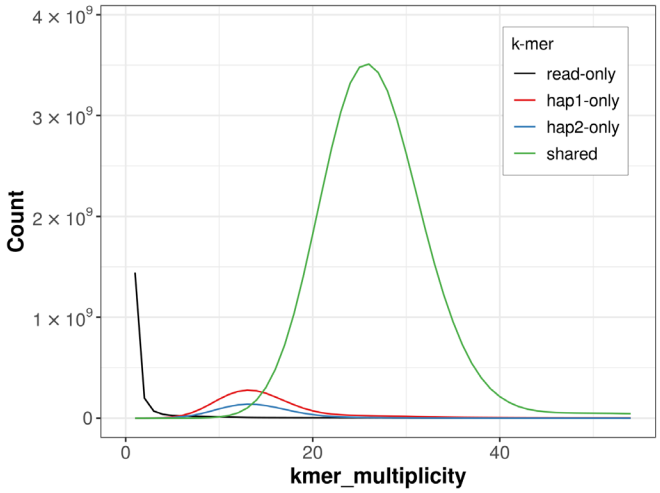

Figure 4. Evaluation of *k*-mer completeness using MerquryFK. This plot illustrates the recovery of *k*-mers from the original read data in the final assemblies. The horizontal axis represents *k*-mer multiplicity, and the vertical axis shows the number of *k*-mers. The black curve represents *k*-mers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

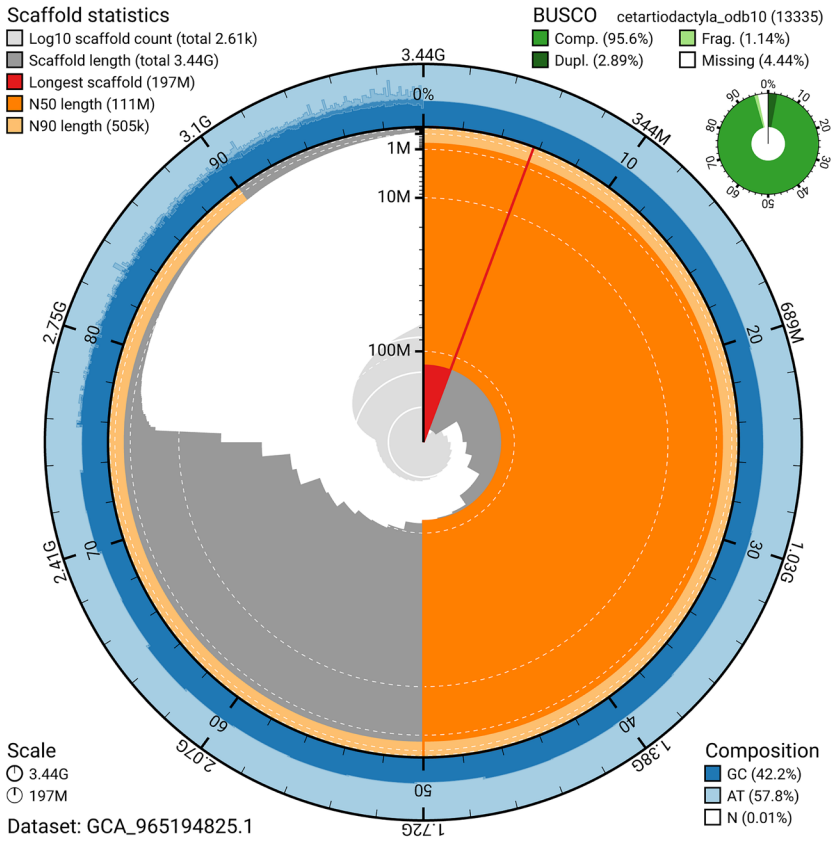

Figure 5. Assembly metrics for mBalPhy2.hap1.1. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1 000 bins around the circumference. The outermost blue 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.

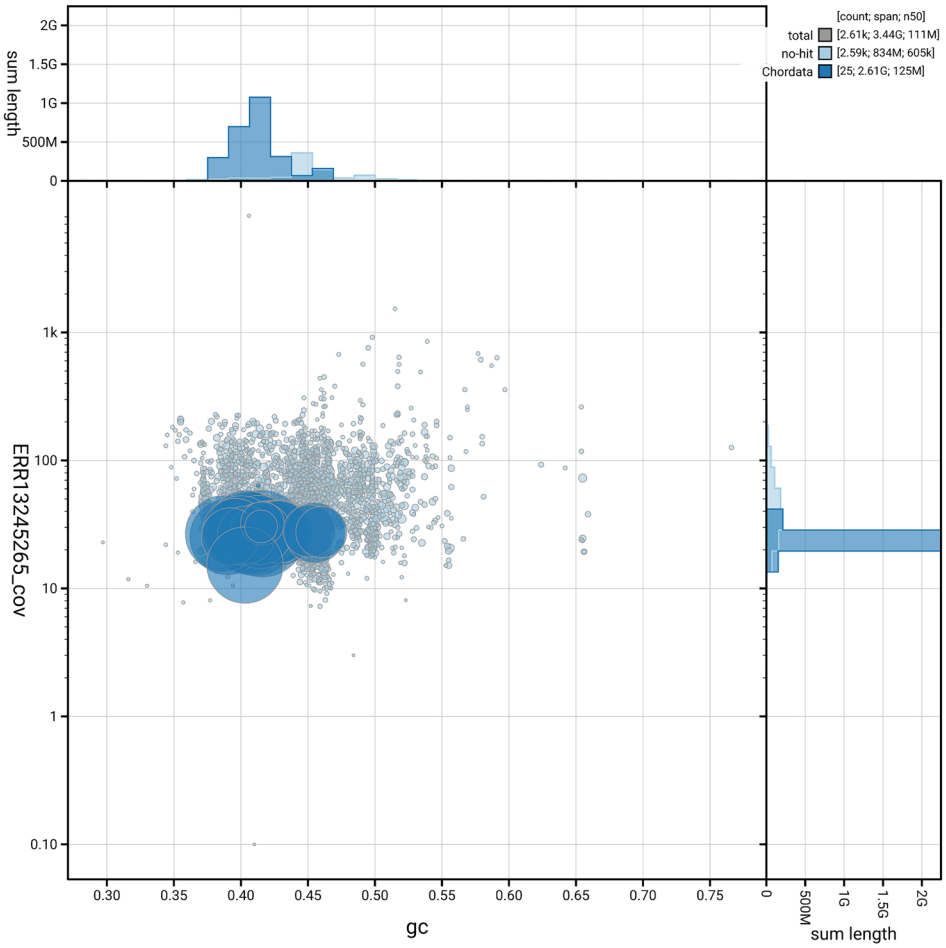

Figure 6. BlobToolKit GC-coverage plot for mBalPhy2.hap1.1. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. The histograms along the axes 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 *Balaenoptera physalus* assembly.

Measure	Value	Benchmark
EBP summary (haplotype 1)	6.8.Q65	6.C.Q40
Contig N50 length	2.08 Mb	≥ 1 Mb
Scaffold N50 length	110.87 Mb	= chromosome N50
Consensus quality (QV)	Haplotype 1: 65.6; haplotype 2: 65.5; combined: 65.6	≥ 40
k-mer completeness	Haplotype 1: 94.59%; Haplotype 2: 89.62%; combined: 99.16%	≥ 95%
BUSCO	C:95.6% [S:92.7%; D:2.9%]; F:1.1%; M:3.3%; n:13 335	S > 90%; D < 5%
Percentage of assembly assigned to chromosomes	79.11%	≥ 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: Balaenoptera physalus (fin whale). Accession number PRJEB76339. The genome sequence is released openly for reuse. The *Balaenoptera physalus* genome sequencing initiative is part of the Darwin Tree of Life Project (PRJEB40665), the Sanger Institute Tree of Life Programme (PRJEB43745), the Cetacean Genomes Project (PRJNA1020146) and the 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.

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	4.4.5	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	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	v0.7.1	https://github.com/sanger-tol/blobtoolkit
sanger-tol/curationpretext	1.4.2	https://github.com/sanger-tol/curationpretext

Software	Version	Source
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

Aguilar A, García-Vernet R: **Fin whale: *Balaenoptera physalus* (Linnaeus, 1758).** In: B. Würsig, J. G. M. Thewissen, and K. M. Kovacs (eds), *Encyclopedia of Marine Mammals.* San Diego: Academic Press, 2018; 368–71.

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

Archer FI, Brownell RL Jr, Hancock-Hanser BL, *et al.*: **Revision of fin whale *Balaenoptera physalus* (Linnaeus, 1758) subspecies using genetics.** *J Mammal.* 2019; **100**(5): 1653–70.
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, 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

Cooke JG: **Balaenoptera physalus.** The IUCN red list of threatened species 2018: e.T2478A50349982. 2018.
Publisher 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

Edwards EF, Hall C, Moore TJ, *et al.*: **Global distribution of fin whales *Balaenoptera physalus* in the post-whaling era (1980–2012).** *Mammal Review.* 2015; **45**(4): 197–214.
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

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

Nigenda-Morales SF, Lin MX, Nuñez-Valencia PG, *et al.*: **The genomic footprint of whaling and isolation in fin whale populations.** *Nat Commun.* 2023; **14**(1): 5465.
PubMed Abstract | Publisher Full Text | Free Full Text

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

Rocha RCJ, Clapham PJ, Ivashchenko YV: **Emptying the Oceans: a summary of industrial whaling catches in the 20th century.** *Marine Fisheries Review.* 2014; **76**(4): 1–18.
Publisher 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 05 February 2026

https://doi.org/10.21956/wellcomeopenres.27998.r144117

© **2026 Castro L.** 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.

✔ **L Filipe C Castro**
University of Porto, Porto, Portugal

Great dataset for a key cetacea species. The methods and description are mostly appropriate.

minor details missing:

"The sex chromosomes were identified by the depth of PacBio read coverage." - perhaps some more detail for validation?

"The genome will be annotated using available RNA-Seq data" - not clear what RNA seq data is (will be) 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?
Partly

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

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Comparative Genomics, Evolution and Physiology

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 24 January 2026

https://doi.org/10.21956/wellcomeopenres.27998.r144114

© 2026 Nery M. 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.

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

Great to know that the fin whale genome is ready. This manuscript presents a high-quality, chromosomal-level genome assembly for *Balaenoptera physalus* (fin whale) as part of the Darwin Tree of Life Project. The work is technically sound, and the assembly meets most standard quality benchmarks.

My only concern relates to sex chromosome identification. The Methods state that sex chromosomes were identified "by the depth of PacBio read coverage," but the specific criteria used to distinguish them are not described. It is unclear, for example, what coverage thresholds or comparative expectations (e.g., autosomes vs. sex chromosomes) were applied. In addition, I do not readily observe coverage differences in the Hi-C contact map, although I acknowledge that this may reflect my limited expertise with Hi-C–based sex chromosome inference.

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.