



DATA NOTE

The genome sequence of the harbour porpoise, *Phocoena*

phocoena (Linnaeus, 1758)

[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

¹Scottish Marine Animal Stranding Scheme, University of Glasgow Institute of Biodiversity Animal Health and Comparative Medicine, Glasgow, Scotland, UK

²NOAA Fisheries Southwest Fisheries Science Center, La Jolla, California, USA

V1 First published: 08 Apr 2025, 10:181
<https://doi.org/10.12688/wellcomeopenres.24011.1>

Latest published: 08 Apr 2025, 10:181
<https://doi.org/10.12688/wellcomeopenres.24011.1>

Abstract

We present a genome assembly from a female specimen of *Phocoena phocoena* (harbour porpoise; Chordata; Mammalia; Artiodactyla; Phocoenidae). The genome sequence has a total length of 2,512.71 megabases. Most of the assembly (94.41%) is scaffolded into 22 chromosomal pseudomolecules, including the X sex chromosome. The mitochondrial genome has also been assembled, with a length of 16.38 kilobases.

Keywords

Phocoena phocoena, harbour porpoise, genome sequence, chromosomal, Artiodactyla

Open Peer Review

Approval Status  

	1	2
version 1 08 Apr 2025	 view	 view

1. **Takushi Kishida** , Nihon University,
Kanagawa, Japan

2. **Mariana F. Nery** , Universidade Estadual
de Campinas, Campinas, Brazil

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



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

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

Author roles: Davison NJ: Investigation, Resources; Morin PA: Investigation, Resources, 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 the harbour porpoise, *Phocoena phocoena* (Linnaeus, 1758) [version 1; peer review: 2 approved]

Wellcome Open Research 2025, 10:181 <https://doi.org/10.12688/wellcomeopenres.24011.1>

First published: 08 Apr 2025, 10:181 <https://doi.org/10.12688/wellcomeopenres.24011.1>

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; Phocoenidae; *Phocoena*; *Phocoena phocoena* (Linnaeus, 1758) (NCBI:txid9742)

Background

The porpoise family consists of seven recognised species, including the smallest and most endangered cetacean, the vaquita (*Phocoena sinus*). In contrast to the vaquita, the congeneric harbour porpoise (*Phocoena phocoena*) is abundant and widely distributed in temperate and boreal waters of the Northern hemisphere, primarily in coastal waters, divided among subspecies in the North Pacific (*P. p. vomerina*), North Atlantic (*P. p. phocoena*), and Black Sea (*P. p. relicta*) (Society for Marine Mammalogy Committee on Taxonomy, 2024). There are at least two other potential unnamed subspecies, in the western North Pacific (Rice, 1998) and eastern Atlantic off the Iberian Peninsula and northern Africa (Fontaine, 2016; Fontaine *et al.*, 2014). Within subspecies, population structure has been tied to isolation by distance and variation in habitat, with some evidence of local adaptation (Ben Chehida *et al.*, 2021; Morin *et al.*, 2021).

Harbour porpoises are considered to be abundant (Bjorge & Tolley, 2018) and listed as “Least Concern” by the IUCN globally, but are classified as critically endangered and endangered in the Baltic Sea and Black Sea, respectively (IUCNredlist.org, consulted 29 January 2025). Distribution in coastal habitat exposes harbour porpoises to regional threats including bycatch in fishing gear, hunting (Greenland), increasing vessel traffic and pollution, eutrophication, offshore windmill parks, and climate change (Bjorge & Tolley, 2018).

The genome of the harbour porpoise, *Phocoena phocoena*, 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 chromosome-level genome sequence for *Phocoena phocoena*, based on a specimen from Roseisle, Moray, Scotland (Figure 1).

Genome sequence report

Sequencing data

The genome of a specimen of *Phocoena phocoena* (Figure 1) was sequenced using Pacific Biosciences single-molecule HiFi long reads, generating 63.11 Gb (gigabases) from 7.10 million reads. GenomeScope analysis of the PacBio HiFi data estimated the haploid genome size at 2,542 Mb. Based on this estimated genome size, the sequencing data provided approximately 24.0x coverage of the genome. Chromosome conformation Hi-C sequencing produced 444.98 Gb from 2,946.86 million reads. Table 1 summarises the specimen and sequencing information.



Figure 1. Photograph of the *Phocoena phocoena* (mPhoPho1) specimen used for genome sequencing.

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 19 misjoins or missing joins and removed two haplotypic duplications. These interventions decreased the scaffold count by 1.79% and increased the scaffold N50 by 16.06%. The final assembly has a total length of 2,512.71 Mb in 439 scaffolds, with 966 gaps, and a scaffold N50 of 110.19 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 (94.41%) was assigned to 22 chromosomal-level scaffolds, representing 21 autosomes and the X sex chromosome. These chromosome-level scaffolds, confirmed by Hi-C data, are named according to size (Figure 5; Table 3). During curation, chromosome X was assigned based on synteny to the genome of *Phocoena sinus* (GCA_008692025.1).

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 62.9. The *k*-mer recovery for the primary

Table 1. Specimen and sequencing data for *Phocoena phocoena*.

Project information			
Study title	Phocoena phocoena (harbor porpoise)		
Umbrella BioProject	PRJEB71542		
Species	<i>Phocoena phocoena</i>		
BioSpecimen	SAMEA111380536		
NCBI taxonomy ID	9742		
Specimen information			
Technology	ToLID	BioSample accession	Organism part
PacBio long read sequencing	mPhoPho1	SAMEA111380544	lung
Hi-C sequencing	mPhoPho1	SAMEA111380544	lung
RNA sequencing	mPhoPho1	SAMEA111380544	lung
Sequencing information			
Platform	Run accession	Read count	Base count (Gb)
Hi-C Illumina NovaSeq 6000	ERR12512736	2.95e+09	444.98
PacBio Revio	ERR12408792	4.54e+06	38.21
PacBio Sequel IIe	ERR12408793	2.55e+06	24.91
RNA Illumina NovaSeq 6000	ERR12512737	4.13e+07	6.24

haplotype is 94.81%, and for the alternate haplotype 94.23%; the combined primary and alternate assemblies have a *k*-mer recovery of 99.70%. BUSCO analysis using the cetartiodactyla_odb10 reference set (*n* = 13,335) identified 95.6% of the expected gene set (single = 94.0%, duplicated = 1.7%).

Table 2 provides assembly metric benchmarks adapted from Rhie *et al.* (2021) and the Earth BioGenome Project Report on Assembly Standards September 2024. The assembly achieves the EBP reference standard of **6.C.Q62**.

Methods

Sample acquisition

A juvenile female *Phocoena phocoena* (specimen ID SAN00002604, ToLID mPhoPho1) was collected from Roseisle, Moray, Scotland (latitude 57.6783, longitude −3.4981) on 2022-02-16. The specimen was collected and identified by Nick Davison (Scottish Marine Animal Stranding Scheme University of Glasgow). 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 preparation and homogenisation, DNA extraction, fragmentation and purification. Detailed protocols are available on protocols.io (Denton *et al.*, 2023b).

The mPhoPho1 sample was prepared for DNA extraction by weighing and dissecting it on dry ice (Jay *et al.*, 2023). Tissue from the lung was homogenised using a PowerMasher II tissue disruptor (Denton *et al.*, 2023a). HMW DNA was extracted using the Automated MagAttract v2 protocol (Oatley *et al.*, 2023a). DNA was sheared into an average fragment size of 12–20 kb in a Megaruptor 3 system (Bates *et al.*, 2023). Sheared DNA was purified by solid-phase reversible immobilisation, using AMPure PB beads to eliminate shorter fragments and concentrate the DNA (Oatley *et al.*, 2023b). 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.

RNA was extracted from lung tissue of mPhoPho1 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

Table 2. Genome assembly data for *Phocoena phocoena*.

Genome assembly		
Assembly name	mPhoPho1.1	
Assembly accession	GCA_963924675.1	
Alternate haplotype accession	GCA_963924725.1	
Assembly level for primary assembly	chromosome	
Span (Mb)	2,512.71	
Number of contigs	1,405	
Number of scaffolds	439	
Longest scaffold (Mb)	187.25	
Assembly metric	Measure	Benchmark
Contig N50 length	3.73 Mb	$\geq 1 \text{ Mb}$
Scaffold N50 length	110.19 Mb	= chromosome N50
Consensus quality (QV)	Primary: 62.7; alternate: 63.1; combined: 62.9	≥ 40
k-mer completeness	Primary: 94.81%; alternate: 94.23%; combined: 99.70%	$\geq 95\%$
BUSCO*	C:95.6%[S:94.0%,D:1.7%], F:0.9%,M:3.5%,n:13,335	$S > 90\%$; $D < 5\%$
Percentage of assembly mapped to chromosomes	94.41%	$\geq 90\%$
Sex chromosomes	X	localised homologous pairs
Organelles	Mitochondrial genome: 16.38 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.

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

Tissue from the lung of the mPhoPho1 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 Diagenode Power Masher-II and BioMasher-II tubes and pestles. Following the Arima-HiC v2 kit manufacturer's instructions, crosslinked 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

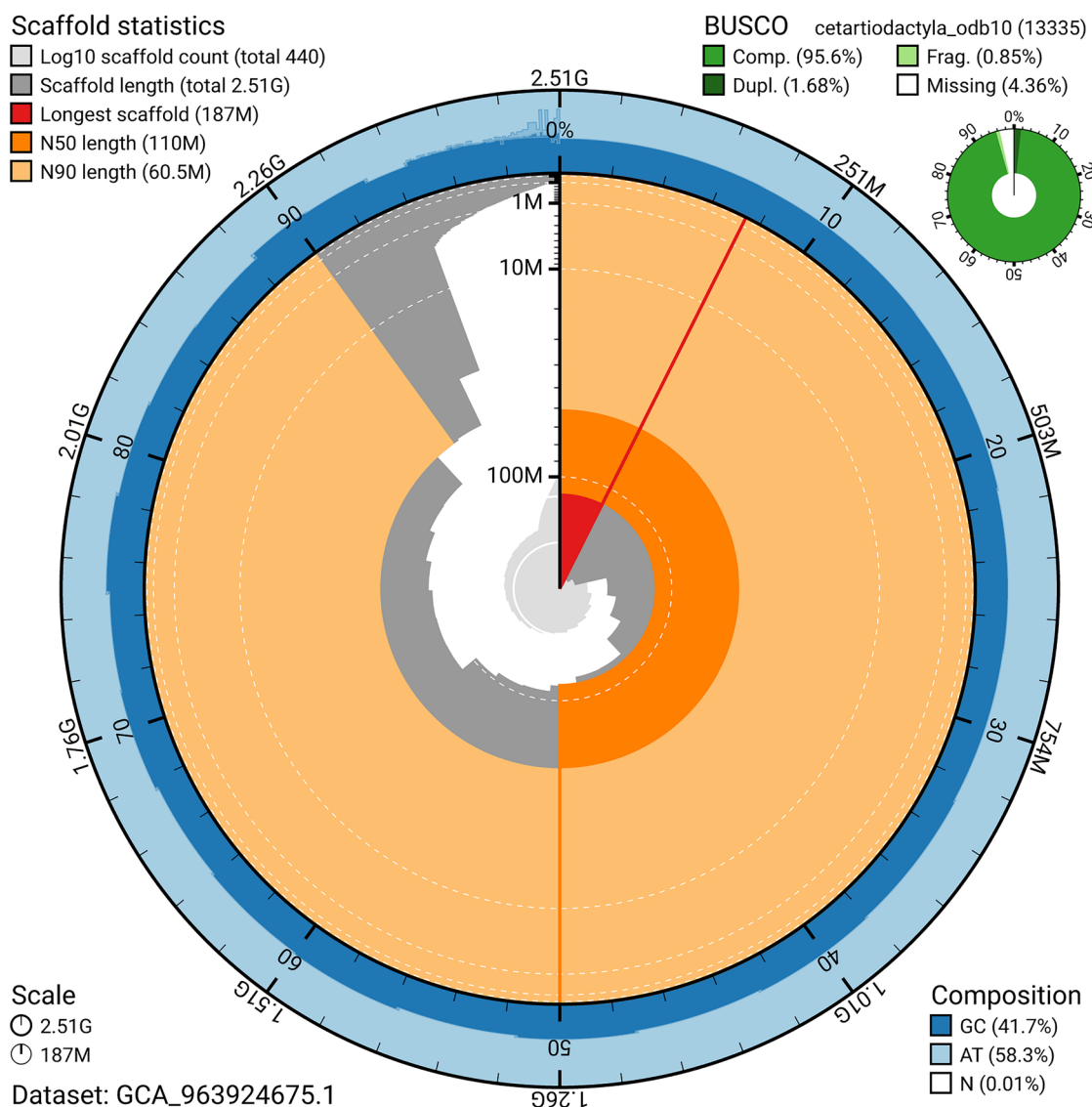


Figure 2. Genome assembly of *Phocoena phocoena*, mPhoPho1.1: metrics. The BlobToolKit snail plot provides an overview of assembly metrics and BUSCO gene completeness. The circumference represents the length of the whole genome sequence, and the main plot is divided into 1,000 bins around the circumference. 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_963924675.1/dataset/GCA_963924675.1/snail.

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

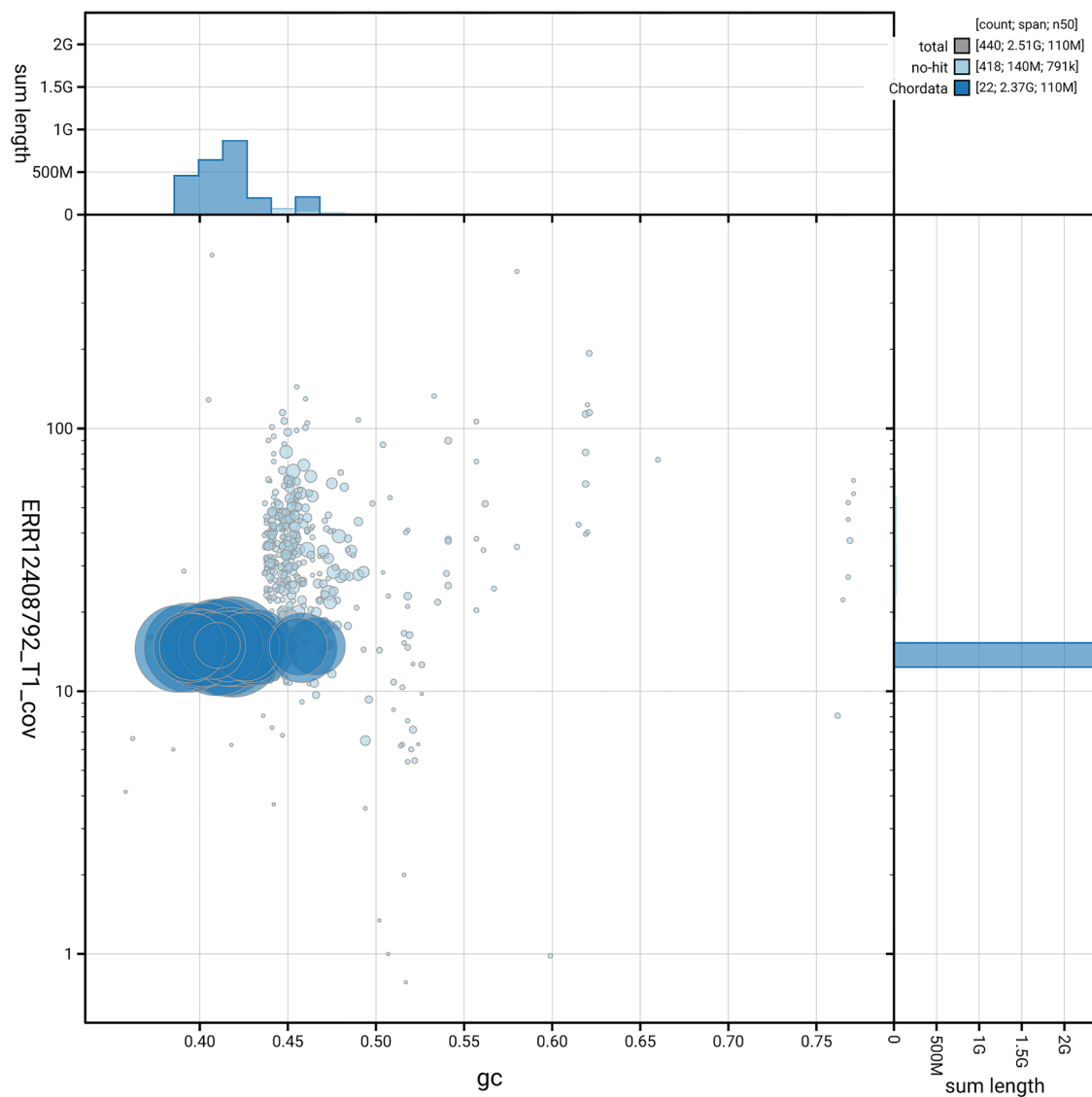


Figure 3. Genome assembly of *Phocoena phocoena*, mPhoPho1.1: BlobToolKit GC-coverage plot. Blob plot showing sequence coverage (vertical axis) and GC content (horizontal axis). The circles represent scaffolds, with the size proportional to scaffold length and the colour representing phylum membership. 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_963924675.1/blob.

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.

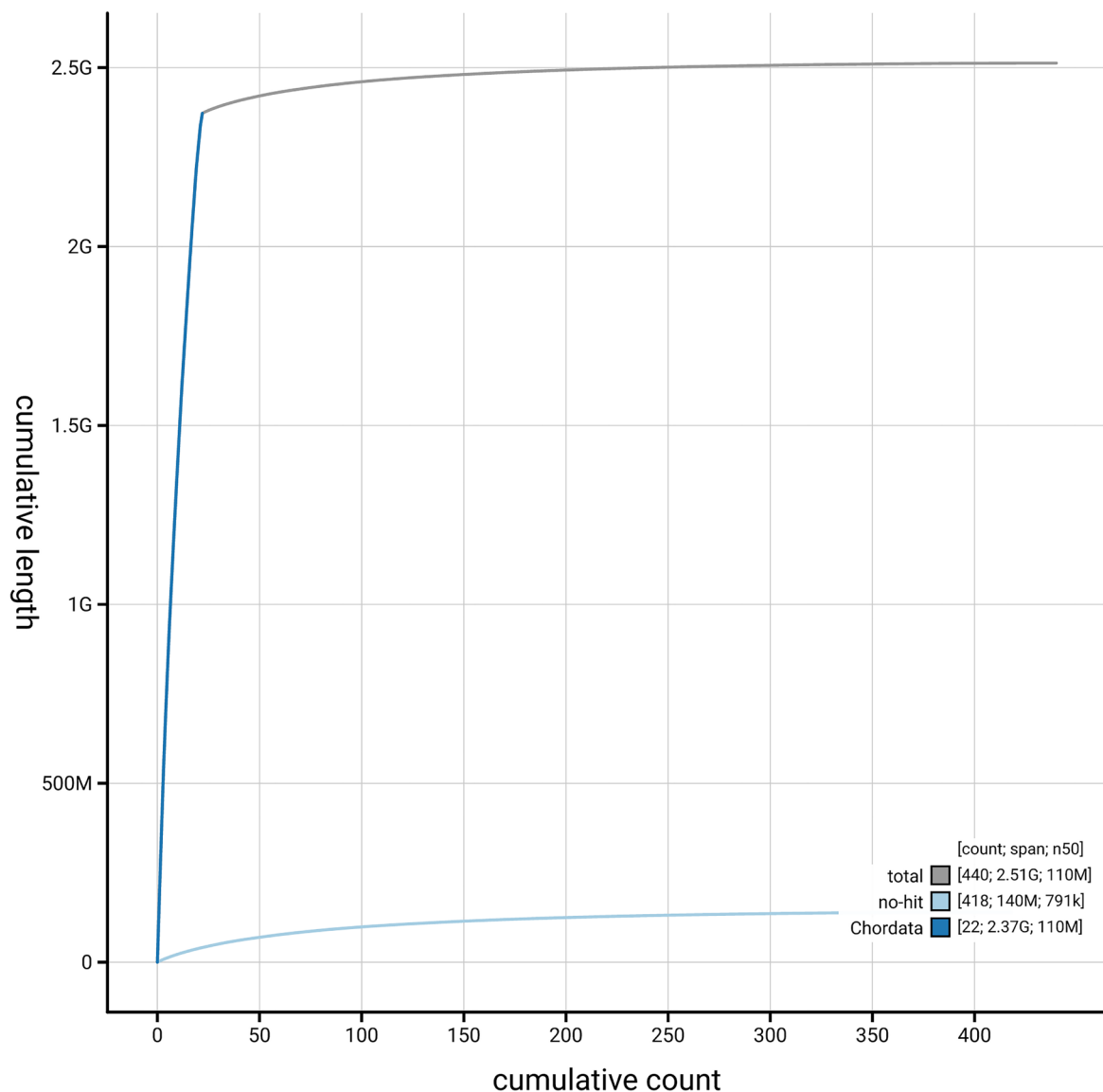


Figure 4. Genome assembly of *Phocoena phocoena*, mPhoPho1.1: BlobToolKit cumulative sequence plot. The grey line shows cumulative length for all scaffolds. Coloured lines show cumulative lengths of scaffolds assigned to each phylum using the buscogenes taxrule. An interactive version of this figure is available at https://blobtoolkit.genomehubs.org/view/GCA_963924675.1/dataset/GCA_963924675.1/cumulative.

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 were mapped to the primary contigs using bwa-mem2 (Vasimuddin *et al.*, 2019). The

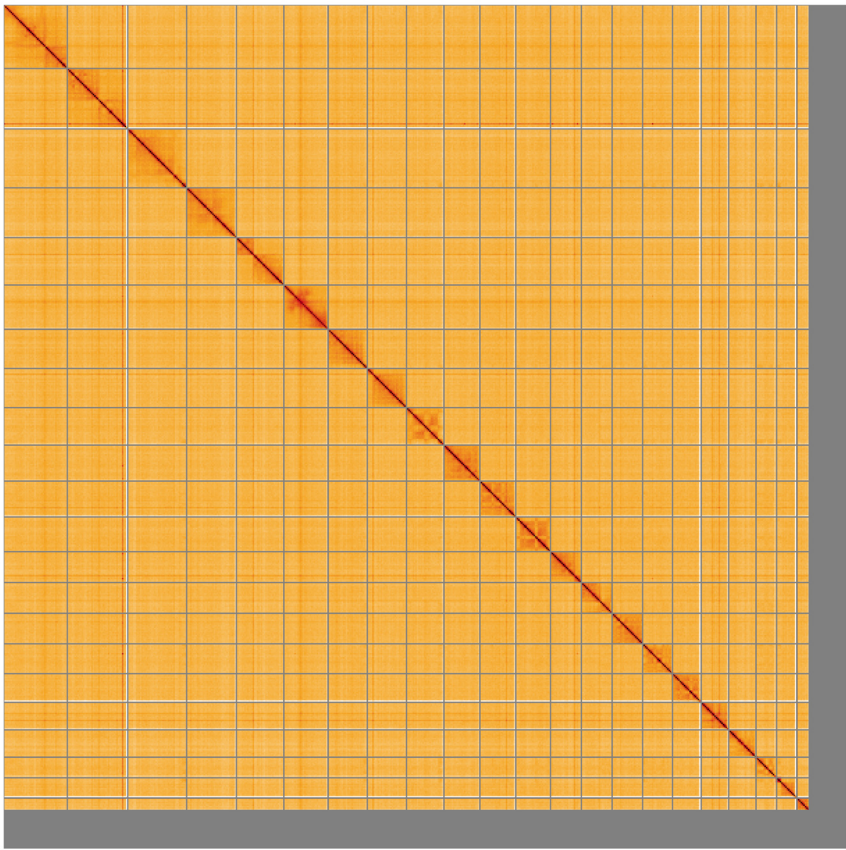


Figure 5. Genome assembly of *Phocoena phocoena*: Hi-C contact map of the mPhoPho1.1 assembly, visualised using HiGlass. Chromosomes are shown in order of size from left to right and top to bottom. An interactive version of this figure may be viewed at <https://genome-note-higlass.tol.sanger.ac.uk/l/?d=JCqle2z0Qv6lCoCQ-9Ev8g>.

Table 3. Chromosomal pseudomolecules in the genome assembly of *Phocoena phocoena*, mPhoPho1.

INSDC accession	Name	Length (Mb)	GC%
OZ004757.1	1	187.25	42
OZ004758.1	2	177.8	41.5
OZ004759.1	3	173.52	41
OZ004760.1	4	146.74	39.5
OZ004761.1	5	139.59	39
OZ004763.1	6	115.56	42
OZ004764.1	7	115.38	40
OZ004765.1	8	110.19	42.5
OZ004766.1	9	106.17	40
OZ004767.1	10	105.44	43

INSDC accession	Name	Length (Mb)	GC%
OZ004768.1	11	102.34	41.5
OZ004769.1	12	90.8	39.5
OZ004770.1	13	90.61	43
OZ004771.1	14	90.18	41.5
OZ004772.1	15	87.82	46
OZ004773.1	16	84.4	42.5
OZ004774.1	17	81.04	40.5
OZ004775.1	18	80.65	39.5
OZ004776.1	19	60.5	46.5
OZ004777.1	20	59.63	45.5
OZ004778.1	21	36.43	41
OZ004762.1	X	130.31	40.5
OZ004779.1	MT	0.02	40.5

contigs were further scaffolded using the provided Hi-C data (Rao *et al.*, 2014) 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),

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 (Pointon *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. Sex chromosomes were identified by synteny analysis. 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 ($k = 31$) that were computed prior to genome assembly. The analysis outputs included assembly QV scores and completeness statistics.

A Hi-C contact map was produced for the final version of the assembly. The Hi-C reads were aligned using bwa-mem2 (Vasimuddin *et al.*, 2019) and the alignment files were combined using SAMtools (Danecek *et al.*, 2021). The Hi-C alignments were converted into a contact map using BEDTools (Quinlan & Hall, 2010) and the Cooler tool suite (Abdennur & Mirny, 2020). The contact map was visualised in HiGlass (Kerpedjiev *et al.*, 2018).

The blobtoolkit pipeline is a Nextflow 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 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.

Table 4. Software tools: versions and sources.

Software tool	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	4.3.9	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
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	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	0.19.8-r603	https://github.com/chhyllp123/hifiasm
HiGlass	44086069ee7d4d3f6f3f0012569789ec138f42b84a a44357826c0b6753eb28de	https://github.com/higlass/higlass
MerquryFK	d00d98157618f4e8d1a9190026b19b471055b22e	https://github.com/thegenemyers/MERQURY.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	0.5.1	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

Data availability

European Nucleotide Archive: *Phocoena phocoena* (harbor porpoise). Accession number PRJEB71542; <https://identifiers.org/ena.embl/PRJEB71542>. The genome sequence is released openly for reuse. The *Phocoena phocoena* genome sequencing initiative is part of the Darwin Tree of Life (DTOL)

project, 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](#).

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

- Abdennur N, Mirny LA: **Cooler: scalable storage for Hi-C data and other genomically labeled arrays.** *Bioinformatics.* 2020; **36**(1): 311–316. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- 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](#)
- Bates A, Clayton-Lucey I, Howard C: **Sanger Tree of Life HMW DNA fragmentation: diagenode Megaruptor[®]3 for LI PacBio.** *protocols.io.* 2023. [Publisher Full Text](#)
- Ben Chehida Y, Loughnane R, Thumlop J, et al.: **No leading-edge effect in North Atlantic harbor porpoises: evolutionary and conservation implications.** *Evol Appl.* 2021; **14**(6): 1588–1611. [PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Björge A, Tolley KA: **Harbor porpoise: *Phocoena phocoena*.** In: Würsig B., Thewissen J. G. M., and Kovacs, K. M. (eds.) *Encyclopedia of Marine Mammals*. 3rd ed. San Diego: Academic Press, 2018; 448–451. [Publisher 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, Grünig 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](#)
- 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](#)
- Denton A, Oatley G, Cornwell C, et al.: **Sanger Tree of Life sample homogenisation: PowerMash.** *protocols.io.* 2023a. [Publisher Full Text](#)
- Denton A, Yatsenko H, Jay J, et al.: **Sanger Tree of Life wet laboratory protocol collection V.1.** *protocols.io.* 2023b. [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, Denton A, Yatsenko H, 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](#)
- Fontaine MC: **Harbour Porpoises, *Phocoena phocoena*, in the Mediterranean Sea and adjacent regions: biogeographic relicts of the last glacial period.** *Adv Mar Biol.* 2016; **75**: 333–358. [PubMed Abstract](#) | [Publisher Full Text](#)
- Fontaine MC, Roland K, Calves I, et al.: **Postglacial climate changes and rise of three ecotypes of harbour porpoises, *Phocoena phocoena*, in western Palearctic waters.** *Mol Ecol.* 2014; **23**(13): 3306–21. [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](#)
- Grüning 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): g1aa153. [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](#)
- Lawnczak 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, Seppely 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](#)
- Morin PA, Forester BR, Forney KA, et al.: **Population structure in a continuously distributed coastal marine species, the harbor porpoise,**

based on microhaplotypes derived from poor-quality samples. *Mol Ecol*. 2021; **30**(6): 1457–1476.

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

Oatley G, Denton A, Howard C: **Sanger Tree of Life HMW DNA extraction: automated MagAttract v.2.** *protocols.io*. 2023a.

[Publisher Full Text](#)

Oatley G, Sampaio F, Howard C: **Sanger Tree of Life fragmented DNA clean up: automated SPRI.** *protocols.io*. 2023b.

[Publisher Full Text](#)

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

[Publisher Full Text](#)

Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics*. 2010; **26**(6): 841–842.

[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.*: **Merquy: reference-free quality, completeness, and phasing assessment for genome assemblies.** *Genome Biol*. 2020; **21**(1): 245.

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

Rice DW: **Marine mammals of the world. Systematics and distribution.** 1998; **4**.

[Reference Source](#)

Society for Marine Mammalogy Committee on Taxonomy: **List of marine mammal species and subspecies.** 2024; [Accessed 30 October 2024].

[Reference Source](#)

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](#)

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 17 April 2025

<https://doi.org/10.21956/wellcomeopenres.26491.r122161>

© 2025 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 

Universidade Estadual de Campinas, Campinas, State of São Paulo, Brazil

The manuscript presents a high-quality, chromosome-level genome assembly for the harbour porpoise (*Phocoena phocoena*), an odontocete cetacean broadly distributed in temperate coastal waters of the Northern Hemisphere. The assembly is highly contiguous and complete, with 94.41% of the sequence assigned to 22 chromosomal pseudomolecules. The authors employed PacBio HiFi long-read sequencing combined with Hi-C data, resulting in excellent quality metrics. The methodology is robust and well-documented, including protocols for high molecular weight DNA and RNA extraction, library preparation, and genome assembly. The genome meets or exceeds current standards established by the Earth BioGenome Project. This genome assembly represents a valuable resource for marine mammal genomics and conservation biology. It is well-executed, well-described, and openly accessible. Minor point: while RNA-Seq data is mentioned, its role in gene annotation or transcriptome analysis is not discussed.

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, molecular evolution

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

Reviewer Report 17 April 2025

<https://doi.org/10.21956/wellcomeopenres.26491.r122159>

© 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 harbor porpoise 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.
