



OPEN

DATA DESCRIPTOR

Novel *Megaptera novaeangliae* (Humpback whale) haplotype chromosome-level reference genome

Maria-Vittoria Carminati^{1,2}✉, Vlonjat Lonnie Gashi¹, Ruiqi Li¹, Daniel Jacob Klee¹, Sara Rose Padula¹, Ajay Manish Patel¹, Andy Dick Yee Tan¹, Jacqueline Mattos^{1,3} & Nolan Kane¹

The sequencing of a kidney sample (KW2013002) from a stranded *Megaptera novaeangliae* (Humpback whale) calf is the first chromosome-level reference genome for this species¹. The calf, a 457 cm and 2,500 lbs male, was found stranded in Hawai'i Kai, HI, in 2013 and was marked as abandoned/orphaned. In 2023, 1 g of kidney was sequenced with PacBio long-read DNA sequencing, chromatin conformation capture (Hi-C), RNA sequencing, and mitochondrial sequencing to comprehensively characterize the genome and transcriptome of *M. novaeangliae*. Data validation includes a synteny analysis, mitochondrial annotation, and a comparison of BUSCO scores (scaffold v. reference genome and *Balaenoptera musculus* (Blue whale) v. *M. novaeangliae*). BUSCO analysis was performed on an *M. novaeangliae* scaffold-level assembly to determine genomic completeness of the reference genome, with a scaffold BUSCO score of 91.2% versus a score of 95.4%. Synteny analysis was performed using the *B. musculus* genome as comparison to determine chromosome-level coverage and structure. Further, a time-based phylogenetic tree was constructed using the sequenced data and publicly available genomes.

Background & Summary

M. novaeangliae are significant marine organisms with a broad geographical distribution, inhabiting all major oceans from the polar regions to the tropics². These whales undertake extensive migrations, traveling thousands of miles annually between feeding grounds in high-latitude waters during the summer and breeding grounds in warmer, low-latitude waters during the winter². Their role in marine ecosystems is critical, as they contribute to nutrient cycling through the release of fecal plumes that stimulate the growth of phytoplankton, forming the base of the oceanic food web^{3,4}. Humpback whales are known for their complex vocalizations and acrobatic behaviors, which have made them a focal point for marine tourism and research^{2,3}. Their populations have experienced significant recovery following historical whaling pressures, highlighting the impact of conservation efforts. However, they continue to face threats from entanglement in fishing gear, ship strikes, and environmental changes, necessitating ongoing monitoring and protective measures.

M. novaeangliae, known for its cosmopolitan distribution, has historically been classified as a single species. However, recent analyses suggest the presence of several subspecies with distinct evolutionary trajectories². Despite this, research into *M. novaeangliae* genetics has been limited by the absence of a complete reference genome.

Although previous efforts have yielded a scaffold-level assembly of the *M. novaeangliae* genome, this falls short of the “platinum-level” standard set by the Cetacean Genome Project (CGP) in 2020^{5–7}. The scaffold-level assembly was not a chromosome-level assembly and contained gaps^{6,7}. Our study aimed to bridge this gap by generating a high-quality, chromosome-level assembly of the *M. novaeangliae* genome^{1,8,9}. This reference genome not only enhances our understanding of *M. novaeangliae* subspeciation but also provides valuable insights into its genetic variation, underlying cell regulation, and the evolutionary mechanisms driving its large size and cancer resistance¹.

¹Department of Ecology and Evolutionary Biology, University of Colorado Boulder, Boulder, CO, USA. ²The Unknown Scientist Institute, Aurora, CO, USA. ³Laboratory of Evolutionary Ecology and Plant Genomics, Department of Plant Biology, Institute of Biology, University of Campinas, São Paulo, Brazil. ✉e-mail: Maria-Vittoria.Carminati@colorado.edu

Using the CGP's platinum-standard criteria, we assembled the *M. novaeangliae* genome at the chromosome level, with one sequence per chromosome in haplotype format^{1,8,9}. This comprehensive assembly allowed us to conduct a range of analyses, including mtDNA annotation, a synteny analysis, generating a phylogenetic tree, comparing the genome against mammalian BUSCO databases, and gene duplication investigations. The availability of this chromosome-level reference genome will catalyze future research on *M. novaeangliae* and cetaceans, offering unprecedented opportunities to explore their genetic diversity, evolution, and conservation.

Methods

Sample information. A kidney sample (KW2013002) was collected from a *M. novaeangliae* calf on January 15, 2013, in Hawai'i Kai, HI, and deposited at the National Institutes of Standards and Technology (NIST). The sample was not collected by the authors so information regarding collection is limited to that presented herein. The calf, a 457 cm and 2,500 lbs male at the time of necropsy, was first observed on January 14, 2013, in shallow water and died between January 14 and January 15, 2013, via stranding. The calf was marked as abandoned/orphaned. In 2023, 1 g of KW2013002 was sampled for sequencing by Cantata Bio.

PacBio long reads DNA sequencing. Quantification of DNA samples was performed using the Qubit 2.0 Fluorometer. For the construction of the PacBio SMRTbell library, targeting an insert size of approximately 20 kb, the SMRTbell Express Template Prep Kit 2.0 was employed following the manufacturer's recommended protocol and default settings. The library was subsequently prepared for sequencing by binding to polymerase using the Sequel II Binding Kit 2.0 (PacBio) and loaded onto the PacBio Sequel II system. Sequencing was executed using PacBio Sequel II 8 M SMRT cells to ensure comprehensive coverage and high-quality reads.

Quality control of the extracted DNA was performed using nanodrop and gel. The OmniC library quality control was done using the Hifiasm draft assembly and showed a high amount of long-range linkage reads. The OmniC sequencing data was also quality controlled to examine Q30%, and the quality score matched the Illumina standard. The scaffolding algorithm HiRise also has a built-in quality control that uses only reads with a map score of over 40.

Chromatin was fixed *in situ* within the nucleus using formaldehyde, followed by digestion with DNase I. The processed chromatin had its ends repaired and was then ligated to a biotinylated bridge adapter, facilitating proximity ligation of adapter-containing ends. Post-proximity ligation, the crosslinks were reversed, and the DNA was purified—a critical step involved treating the purified DNA to eliminate any non-internal biotin. The sequencing libraries were prepared using NEBNext Ultra enzymes and Illumina-compatible adapters, with biotin-containing fragments isolated using streptavidin beads before PCR enrichment. Sequencing was performed on an Illumina HiSeqX platform to achieve approximately 30x coverage.

Contig assembling and scaffolding. The *de novo* assembly process utilized PacBio CCS reads and Omni-C reads as input for HiC-Hifiasm, employing default parameters. This approach facilitated the generation of a separate *de novo* assembly for each haplotype, enhancing the accuracy and integrity of the genomic reconstruction.

The scaffolding phase involved the integration of the *de novo* assembly with Dovetail Omni-C library reads through HiRise, a software pipeline tailored for scaffolding genome assemblies using proximity ligation data. Alignment of Omni-C library sequences to the draft assembly was achieved using bwa, with the mapped read pairs analyzed by HiRise to construct a likelihood model for genomic distance. This model, along with additional information from the synteny analysis (see below), informed the identification and correction of misjoins, the scoring of potential joins, and the execution of joins exceeding a defined confidence threshold.

Synteny analysis. The *M. novaeangliae* newly-assembled scaffolds were mapped to the *B. musculus* whole genome (GenBank GCA_009873245.3) in order to map the synteny between the two species^{1,8,10}. A synteny analysis was performed using JupiterPlot 1.0^{11–13}, a software tool that uses circos-based consistency plots to map a given set of scaffolds with a reference genome.

RNA sequencing. Total RNA was extracted employing the QIAGEN RNeasy Plus Kit, adhering to the manufacturer's instructions. Quantification of RNA involved the Qubit RNA Assay and the TapeStation 4200 system. Before library preparation, DNase treatment was applied, followed by AMPure bead cleanup and rRNA depletion using QIAGEN FastSelect -HMR. The NEBNext Ultra II RNA Library Prep Kit was used for library preparation per the manufacturer's protocols. Sequencing of the prepared libraries was conducted on the NovaSeq 6000 platform, utilizing a 2 × 150 bp configuration to ensure comprehensive transcriptome coverage.

Data Records

The RNA-Seq (SRX24476266), Omni-C (SRX24476265), and raw reads of the *M. novaeangliae* genome (SRX24476264) are available on the NCBI website under SRA SRP506011⁹. The mitochondrial sequence is publicly available at GenBank with accession number PP475430.1¹⁴. The chromosome-level genome assembly is available via GenBank at NCBI with submission number SUB14625650 and accession number JBGMDX000000000 (WGS JBGMDX010000001:JBGMDX010001355)¹. The raw genome assembly is publicly available via Dryad⁸.

Technical Validation

The quality of the genome assembly was evaluated using a comprehensive set of metrics and analyses. The scaffold assembly had a total length of 2.27 Mbp and an N50 of 9.15 Mbp^{1,6,7}. The reference input assembly had a total length of 3.15 Gbp with an N50 of 36.344 Mbp, indicating a high degree of contiguity⁹. The reference genome improved upon this, with a total length of 2.48 Gbp and a significantly higher N50 of 116.74 Mbp, suggesting a substantial increase in contiguity and completeness¹. The N90 values also improved from 2.35 Mbp to 83.01

	Mammalian BUSCO genes			
	Single Copy Complete	Duplicated	Fragmented	Missing
<i>M. novaeangliae</i>	8,555	258	127	286
Percentage of mammalian BUSCO genes	92.70%	2.80%	1.40%	3.10%
Shared between <i>M. novaeangliae</i> and <i>B. musculus</i>	8,330	162	83	211
Percentage of <i>M. novaeangliae</i> genes shared with <i>B. musculus</i>	97.30%	62.80%	65.30%	73.80%

Table 1. Summary of assembly characteristics for *M. novaeangliae* reference genomes, comparing the reference genome and the scaffold.

<i>M. novaeangliae</i> assemblies	# of scaffolds	Total length (Gb)	Longest sequence (Mb)	N50 (Mb)	N90 (Mb)	L50	L90	Missing data (% of total length)
Reference	23	2.48	191.56	116.74	83.01	9	19	0.0003
Scaffold	2558	2.27	29.35	9.14	2.35	79	273	5.4534

Table 2. Summary of the mammalian BUSCO analysis on the *M. novaeangliae* reference genome and a comparison between it and the closely related *B. musculus* reference genome.

Mbp, indicating that a larger proportion of the genome is represented in longer scaffolds, further improving contiguity¹. The largest scaffold for the reference genome was 191,558,633 bp, which also indicates enhanced contiguity achieved through the HiRise scaffolding process¹. Additionally, while the input assembly had 0 gaps, the reference genome had 59 gaps, which could potentially represent regions of the genome that are difficult to assemble or highly repetitive.

Furthermore, the evaluation included an assessment of genome completeness using BUSCOs, which are conserved single-copy genes. The reference genome showed high completeness, with 98.04% of BUSCOs being complete, and 96.08% being complete and single copy. This indicates that the assembly is highly representative of the expected gene content for eukaryotic species. The use of the Dovetail Omni-C library and the HiRise scaffolding pipeline contributed to the high quality of the assembly by leveraging proximity ligation data to improve contiguity and accuracy. Overall, the genome assembly appears to be of high quality, with improved contiguity and completeness compared to the input assembly.

The resulting comparison shows that the reference genome is a chromosome-level assembly, is longer than the scaffold (2.48 Gbp v. 2.27 Gbp), has longer sequences, and is missing less data (0.0003% v. 5.4534%).

Benchmarking Universal Single-Copy Orthologs (BUSCO) Gene Analysis Against Mammalian Genomes. BUSCO analysis was run with the mammalian ortholog dataset to compare the assembly against a set of highly conserved genes present in placental mammals. When compared to the existing scaffold, the reference genome had higher numbers of conserved BUSCO scores specific to placental mammals:

1. *M. novaeangliae* Reference Genome
- Complete BUSCOs (C): 95.4% [Single-copy: 92.7%, Duplicated: 2.7%]
 - Fragmented BUSCOs (F): 1.0%
 - Missing BUSCOs (M): 3.6%
 - Total number of BUSCOs searched (n): 12,234
2. *M. novaeangliae* Scaffold Genome
- Complete BUSCOs (C): 91.2% [Duplicated: 0.8%]
 - Fragmented BUSCOs (F): 4.8%
 - Missing BUSCOs (M): 4.0%
 - Total number of BUSCOs searched (n): 6,253

These readings indicate the completeness and duplication status of the genomes being analyzed, with the reference genome showing higher completeness compared to the scaffold genome. The comparison of placental mammal BUSCO scores for reference and scaffold *M. novaeangliae* genome indicates a near 50% reduction in fragmented or missing BUSCO genes (Reference: 4.6%, Scaffold: 8.8%). The percentage of duplicated BUSCOs is also provided, indicating the level of duplication in the identified complete BUSCOs.

To further validate the *M. novaeangliae* reference genome, it was compared to the *B. musculus* genome, giving the following results^{1,8,10}.

1. *B. musculus*
- Complete BUSCOs (C): 95.1% [Single-copy: 92.5%, Duplicated: 2.6%]
 - Fragmented BUSCOs (F): 1.4%

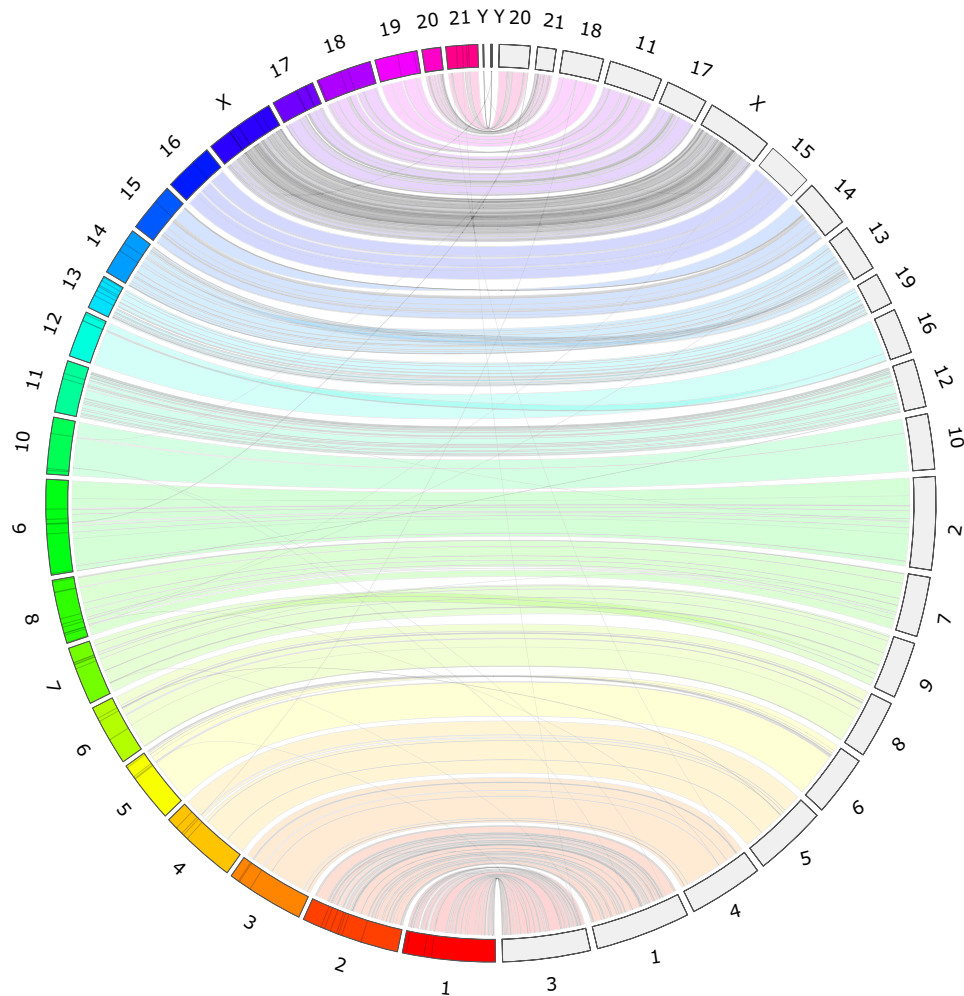


Fig. 1 Jupiterplot of *M. novaeangliae* and *B. musculus* genomes. This figure illustrates the overall alignment of these two genomes, identifying syntenic regions between their chromosomes. Connected regions show the genetic links between both genomes.

- Missing BUSCOs (M): 3.5%
- Total number of BUSCOs searched (n): 9,226

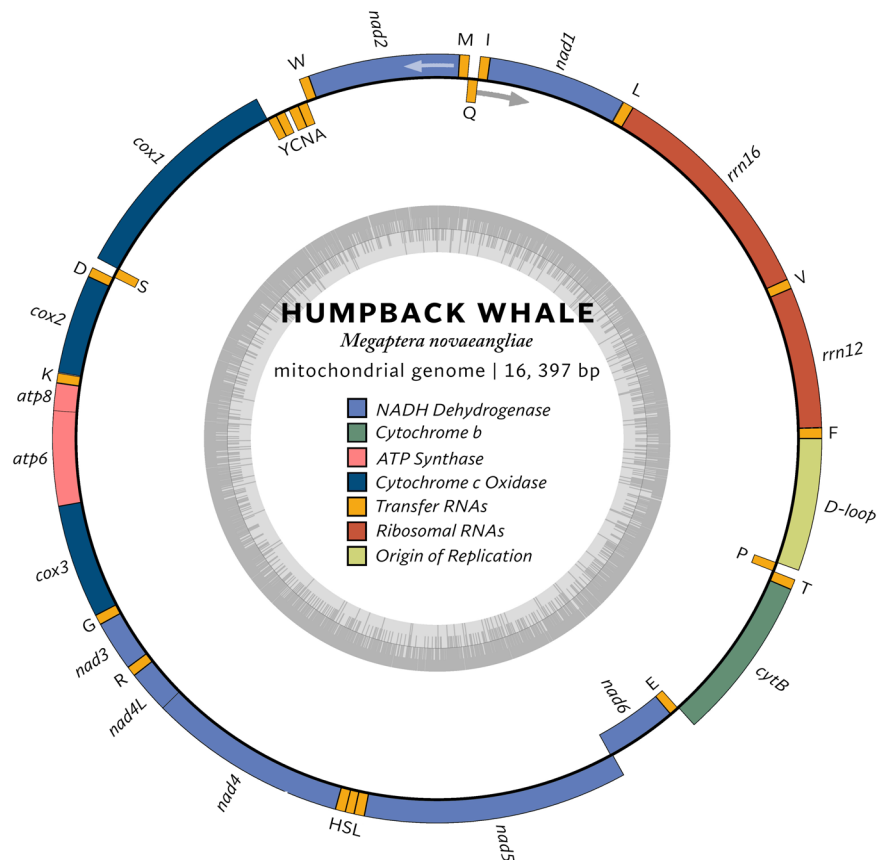
2. *M. novaeangliae*

- Complete BUSCOs (C): 95.4% [Single-copy: 92.7%, Duplicated: 2.7%]
- Fragmented BUSCOs (F): 1.4%
- Missing BUSCOs (M): 3.2%
- Total number of BUSCOs searched (n): 9,226

The comparison indicates that the *M. novaeangliae* reference genome is more complete than the previous scaffold genome and is comparable to the *B. musculus* reference genome (Tables 1, 2).

A comparison within the *M. novaeangliae* and *B. musculus* BUSCO genes was then performed, with the following outcome:

Synteny analysis. Synteny analysis involves comparing the arrangement of genes or genetic elements in different genomes to understand their evolutionary relationships and identify conserved regions. By using Jupiter Plot we identified and visualized these syntenic regions between the humpback whale genome and the blue whale genome. The synteny plot (Fig. 1) shows the syntenic regions between these two genomes, and a high degree of collinearity between them. The X and Y chromosomes are also syntenic, showing a high level of consistency in the evolutionary relationships between these two species.



Mitochondrial DNA analysis. Our study also analyzed the mitochondrial genome of *M. novaeangliae*¹⁴. The complete *M. novaeangliae* genome was sequenced and assembled (Fig. 2). The genome contains the typical set of mitochondrial genes found in vertebrates, including genes coding for ribosomal RNAs, transfer RNAs, and protein-coding genes such as NADH dehydrogenase subunits, cytochrome c oxidase subunits, ATP synthase subunits, and cytochrome b.^{12,13} The genome also contains a D-loop region of 915 bp, which is a non-coding region often used for studying genetic diversity and population structure. Annotation was done by chloroblox/mitos and visualization was done on chlorobox. This sequence represents an important resource for future studies on the evolution, population genetics, and conservation of humpback whales.

Phylogenomic analysis. Protein coding genes were retrieved from 16 mammal genomes downloaded from NCBI (Fig. 3)^{10,15–29}. Along with genes in the *M. novaeangliae* genome, they were clustered into gene families using OrthoFinder v2.27³⁰. The sequences from single copy gene families were aligned with mafft v7.508¹⁰, then trimmed with TrimAL v1.4 (<http://trimal.cgenomics.org/>). The trimmed alignments were concatenated to a super-matrix^{11,31}. IQtree2 was used to infer phylogenetic trees from the super-matrix with 1000 bootstraps³². ModelFinder integrated in the IQtree2 was employed to select the best-fit model. Then the phylogeny was time calibrated using LSD2 embedded in IQtree2 with fossil records and inferred divergent dates (Fig. 3)³³.

The protein-coding proteins were annotated with eggnoG mapper against Gene Ontology (GO), KEGG, EC, CAZy, databases with default settings. 28,109 gene families were identified using OrthoFinder, of which 511 were single copies³⁴. The final concatenated super-matrix was 240,833 bp. The best-fit model is Q.bird + F + I + R6 according to BIC. Among the total of 38660 genes, 21159 (54.73%) were annotated in at least one of the emapper databases. Specifically, 17,574 (45.46%) have gene names, 13,871 (35.88%) feature GO annotations, 3,196 (8.27%) are associated with EC numbers, and 13,461 (34.82%) are linked to KEGG pathways.

Code availability

The data analyses were performed according to the manuals and protocols provided by the developers of the respective bioinformatics tools and no custom code was used in the execution of the study. All software and codes used in this work are publicly accessible, and their corresponding versions are specified in the Methods section.

Received: 21 May 2024; Accepted: 23 September 2024;

Published online: 10 October 2024

References

- Carminati, M.-V. G. *et al.* Megaptera novaeangliae isolate KW2013002, whole genome shotgun sequencing project. *GenBank* <https://identifiers.org/ncbi/insdc:JBGMGX010000001.1> (2024).
- Jackson, J. A. *et al.* Global diversity and oceanic divergence of humpback whales (Megaptera novaeangliae). *Proc. R. Soc. B* **281**, 1786, <https://doi.org/10.1098/rspb.2013.3222> (2014).
- Roman, J. & McCarthy, J. J. The whale pump: marine mammals enhance primary productivity in a coastal basin. *PLoS One* **5**, 10, <https://doi.org/10.1371/journal.pone.0013255> (2010).
- McGowen, M. R. *et al.* Phylogenomic Resolution of the Cetacean Tree of Life Using Target Sequence Capture. *Syst. Biol.* **69**, 479–501, <https://doi.org/10.1093/sysbio/syz068> (2020).
- Morin, P. A. *et al.* Building genomic infrastructure: Sequencing platinum-standard reference-quality genomes of all cetacean species. *Marine Mammal Science* **36**(4), 1356–1366, <https://doi.org/10.1111/mms.12721> (2020).
- GenBank. *Megaptera novaeangliae* (humpback whale) genome assembly, megNov1, GCA_004329385.1. National Center for Biotechnology Information (NCBI). Available from: https://www.ncbi.nlm.nih.gov/assembly/GCA_004329385.1 (2019).
- Tollis, M. *et al.* Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). *Molecular Biology and Evolution* **36**(8), 1746–1763, <https://doi.org/10.1093/molbev/msz099> (2019).
- Carminati, M.-V. G. *et al.* Novel Megaptera novaeangliae (Humpback whale) haplotype reference genome [Dataset]. *Dryad* <https://doi.org/10.5061/dryad.dv41ns271> (2024).
- NCBI Sequence Read Archive <https://www.ncbi.nlm.nih.gov/sra/SRP506011> (2024).
- GenBank. *Balaenoptera musculus* (blue whale) genome assembly, mBalMus1.pri.v3, GCA_009873245.3. National Center for Biotechnology Information (NCBI). Available from: https://www.ncbi.nlm.nih.gov/assembly/GCA_009873245.3 (2020).
- Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics* **20**(4), 1160–1166, <https://doi.org/10.1093/bib/bbx108> (2019).
- Allio, R., Donegà, S., Galtier, N. & Nabholz, B. Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker. *Molecular Biology and Evolution* **34**(11), 2762–2772, <https://doi.org/10.1093/molbev/msx197> (2017).
- Cummins, J. Mitochondrial DNA in mammalian reproduction. *Reviews of reproduction* **3**(3), 172–182, <https://doi.org/10.1530/ror.0.0030172> (1998).
- Carminati, M.-V. G. *et al.* Megaptera novaeangliae voucher NIST KW2013002 mitochondrion, complete genome. <https://identifiers.org/ncbi/nucleotide:PP475430.1> (2024).
- Genome Assembly mEubGla1.1hap2.+XY (*Eubalena glacialis*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_028564815.1/ (2023).
- Genome Assembly mBalAcu1.1 (*Balaenoptera acutorostrata*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_949987535.1/ (2023).
- Genome Assembly mBalRic1.hap2 (*Balaenoptera ricei*) https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_028023285.1/ (2023).
- Genome Assembly mTurTru1.mat.Y (*Tursiops truncatus*) https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_011762595.1/ (2020).
- Genome Assembly mOrcOrc1.1 (*Orcinus orca*) https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_937001465.1/ (2022).
- Genome Assembly mKogBre1hap1type1 (*Kogia breviceps*) https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_026419965.1/ (2022).
- Genome Assembly ASM283717v5 (*Physeter catodon*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_002837175.3/ (2023).
- Genome Assembly Loxafr3.0 (*Loxodonta africana*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000001905.1/ (2009).
- Genome Assembly mHipAmp2.hap2 (*Hippopotamus amphibius kiboko*) https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_030028045.1/ (2023).
- Eschrichtius robustus Genome sequencing and assembly <https://www.ncbi.nlm.nih.gov/bioproject/PRJNA707533/> (2023).

25. Genome Assembly Oros_1.0 (*Odobenus rosmarus divergens*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000321225.1/ (2013).
26. Genome Assembly TriManLat1.0 (*Trichechus manatus latirostris*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000243295.1/ (2012).
27. Genome assembly Neophocaena_asiaeorientalis_V1.1 (*Neophocaena asiaeorientalis asiaeorientalis*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_003031525.2/ (2018).
28. Genome Assembly mGloMel1.1 (*Globicephala melas*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_963455315.1/ (2023).
29. Genome assembly mMesDen1 primary haplotype (*Mesoplodon densirostris*), https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_025265405.1/ (2022).
30. Justin Chu. JupiterPlot. GitHub repository. <https://github.com/JustinChu/JupiterPlot> (2018).
31. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780, <https://doi.org/10.1093/molbev/mst010> (2013).
32. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274, <https://doi.org/10.1093/molbev/msu300> (2015).
33. Smith, S. A. LSD2: Least-Squares Dating for Estimating Species Divergence Times. *Bioinformatics* **35**(21), 4429–4431 (2019).
34. Kuznetsov, D. *et al.* OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res* **51**, D445–D451, <https://doi.org/10.1093/nar/gkac998> (2023).

Acknowledgements

Gratitude to experiment.com, where the funds were raised to fund the sequencing of the reference genome. Many thanks to NIST for providing the *M. novaeangliae* sample and to Cantata Bio for its in-kind support towards the sequencing itself. Tremendous thanks to Dr. Stephane Flibotte at the Life Sciences Institute Bioinformatics Facility, University of British Columbia, for his support processing and uploading the data to GenBank. And of course, special thanks to Justin Viezbicke of the California Stranding Network Coordinator, National Marine Fisheries Service, for overseeing the issuance of a permit that made all of this possible. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund and Prof. Andrew Martin, PhD.

Author contributions

Dr. Maria-Vittoria “Giugi” Carminati: Obtained the parts permit for the sequencing of the DNA, raised funds for the sequencing, assembly, and annotation of the genome, located the sample, coordinated its delivery to Cantata Bio, assembled the team under Dr. Nolan Kane’s supervision to process the genome, attended and contributed to a semester-long whale genomics graduate seminar, provided guidance and feedback for the assembly of validation data, and drafted the manuscript (first draft, revisions, final draft). Dr. Nolan Kane: Led the author team via a semester-long whale genomics seminar, taught the techniques and analysis necessary for data validation, led group discussions, provided reading materials, offered course corrections as needed, and gave feedback. Sara Padula, Ajay Patel, Andy Tan, Vlonjat Lonnie Gashi: Attended the semester-long whale genomics class that contributed to this manuscript, developed the BUSCO scores and comparisons, identified the duplicated, missing, fragmented, and common BUSCO genes, and generated all related figures. Sara Padula also provided additional revisions to the manuscript. Dan Klee: Attended the semester-long whale genomics class that contributed to this manuscript, cleaned up, analyzed, annotated, and uploaded the mitochondrial sequence to GenBank, and reviewed and revised the manuscript portions related to the mitochondrial sequence. Andy Tan and Ruiqi Li: Attended the semester-long whale genomics class that contributed to this manuscript, performed the phylogenomic analysis (both ML IQTree with 1000 bootstraps and time-calibration), wrote the phylogenomics methods section of this manuscript, and generated all phylogeny figures. Andy Tan also generated the circular mitogenome annotation figure in this manuscript. Jacqueline Mattos: Generated the synteny figure and its legend, contributed the methods writeup related to the synteny analysis, and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.-V.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024