A chromosome-level genome assembly for the dugong 
(\textit{Dugong dugon})

Dorothy Nevé Baker\textsuperscript{1,*}, Linelle Abueg\textsuperscript{2}, Merly Escalona\textsuperscript{3}, Katherine A. Farquharson\textsuperscript{4,5}, Janet M. Lanyon\textsuperscript{6}, Diana Le Duc\textsuperscript{7}, Torsten Schöneberg\textsuperscript{8,9}, Dominic Absolon\textsuperscript{10}, Ying Sims\textsuperscript{10}, Olivier Fedrigo\textsuperscript{11}, Erich D. Jarvis\textsuperscript{2,12}, Katherine Belov\textsuperscript{4,5}, Carolyn J. Hogg\textsuperscript{4,5} and Beth Shapiro\textsuperscript{1,12}

\textsuperscript{1}Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, United States, \textsuperscript{2}Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, United States, \textsuperscript{3}Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA, United States, \textsuperscript{4}Faculty of Science, School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW, Australia, \textsuperscript{5}Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science, The University of Sydney, NSW, Australia, \textsuperscript{6}School of Biological Sciences, The University of Queensland, St Lucia, QLD, Australia, \textsuperscript{7}Institute of Human Genetics, University Medical Center Leipzig, Leipzig, Germany, \textsuperscript{8}Medical Faculty, Rudolf Schönheimer Institute of Biochemistry, University of Leipzig, Leipzig, Germany, \textsuperscript{9}School of Medicine, University of Global Health Equity, Kigali, Rwanda, \textsuperscript{10}Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom, \textsuperscript{11}Colossal Biosciences, Cambridge, MA, United States, \textsuperscript{12}Howard Hughes Medical Institute, Chevy Chase, MD, United States

*Corresponding author: Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA, USA. Email: neve.baker@gmail.com

Corresponding Editor: William Murphy

Abstract

The dugong (\textit{Dugong dugon}) is a marine mammal widely distributed throughout the Indo-Pacific and the Red Sea, with a Vulnerable conservation status, and little is known about many of the more peripheral populations, some of which are thought to be close to extinction. We present a de novo high-quality genome assembly for the dugong from an individual belonging to the well-monitored Moreton Bay population in Queensland, Australia. Our assembly uses long-read PacBio HiFi sequencing and Omni-C data following the Vertebrate Genome Project pipeline to reach chromosome-level contiguity (24 chromosome-level scaffolds; 3.16 Gbp) and high completeness (97.9\% complete BUSCOs). We observed relatively high genome-wide heterozygosity, which likely reflects historical population abundance before the last interglacial period, approximately 125,000 yr ago. Demographic inference suggests that dugong populations began declining as sea levels fell after the last interglacial period, likely a result of population fragmentation and habitat loss due to the exposure of seagrass meadows. We find no evidence for ongoing recent inbreeding in this individual. However, runs of homozygosity indicate some past inbreeding. Our draft genome assembly will enable range-wide assessments of genetic diversity and adaptation, facilitate effective management of dugong populations, and allow comparative genomics analyses including with other sirenians, the oldest marine mammal lineage.

Key words: conservation, long-read assembly, marine mammals, sirenians, Vertebrate Genomes Project, whole genome

Introduction

Dugongs (\textit{Dugong dugon}; Fig. 1A) are marine mammals with a broad but fragmented distribution throughout the Indian and western Pacific Oceans (Husar 1978). Dugongs belong to the order Sirenia along with manatees, and are the only extant representative of the family Dugongidae. They are also the closest relative of the Steller’s sea cow, a giant sirenian that was hunted to extinction in the 18th century. Dugongs prefer shallow coastal waters and are mainly herbivorous, relying on seagrass meadows for both food and habitat (Best 1981). Dugongs are a culturally important species to Torres Strait Islander and many coastal Aboriginal communities for cultural ceremonies, hunting, and in custodianship of Sea Country (Leong 1998; Lincoln et al. 2021). Little is published in the literature about dugong behavior—their shy and elusive nature makes them challenging to study in the wild and, unlike many other small marine mammals, they are difficult to maintain in captivity (Bertram and Bertram 1973; Goto et al. 2004). While some areas, such as northern and eastern Australia, have robust ecological monitoring programs for dugongs...
A Chromosome-level Genome Assembly for the Dugong doi:10.1093/jhered/esae003

and co-management programs with Indigenous communities (Tibbetts et al. 2019; Lincoln et al. 2021; Cleger et al. 2023), other dugong populations throughout south Asia and eastern Africa are data deficient (Marsh et al. 2002). The IUCN lists dugongs as Vulnerable; however, some populations are thought to be close to extinction due primarily to habitat destruction and fisheries bycatch (Marsh et al. 1995, 2002). Evidence from aerial surveys, habitat mapping, and interviews with local communities suggests that the global range of dugongs has contracted (Marsh et al. 2002), leaving potentially endangered...
and isolated relict populations—particularly in the western Indian Ocean—and generating concern about loss of genetic diversity (Plön et al. 2019). However, substantial uncertainty remains concerning the global status of dugongs.

Many questions remain relating to dugong demographics, movement, and population structure that can be addressed using whole-genome data. Previous genetic studies have relied primarily on analyzing the distribution of mitochondrial control region haplotypes (Blair et al. 2014; Plön et al. 2019; Srinivas et al. 2020; Garrigue et al. 2022). These studies have shown that dugong mitochondrial haplotypes show significant geographic structure throughout their range and generally high mitochondrial haplotype diversity range-wide (Blair et al. 2014; Seddon et al. 2014; Plön et al. 2019), with lower diversity at the range periphery (Plön et al. 2019; Garrigue et al. 2022). Microsatellite and SNP genotypes also recovered significant geographic structure as well as isolation by distance, reflecting generally low dispersal among dugongs (Seddon et al. 2014; Cope et al. 2015; McGowan et al. 2023). The environmental forces contributing to this structure are not fully understood; however, sea level fluctuations associated with Pleistocene glacial cycles may have allowed range expansion and contraction by repeatedly creating and destroying the shallow near-shore seagrass habitat upon which dugongs rely (Woodruff 2010). For example, much of the marine near-shore environment around northern Australia and southeast Asia—the approximate geographic center of present-day dugong range—was not submerged until the end of the last glacial maximum 17,000 yr ago (Ludt and Rocha 2015). Cryptic marine barriers (e.g. tidal and current patterns) and breaks in seagrass habitat may also play a role (McGowan et al. 2023).

Here, we present a highly contiguous, chromosome-level de novo high-quality genome assembly for the dugong, along with initial estimates of genomic diversity and demographic history. Our assembly provides a resource for future genomic studies of dugong population structure, conservation status, and evolutionary history, and will contribute to the larger Vertebrate Genome Project (Rhee et al. 2021). Along with existing draft-quality genome assemblies for manatees and the extinct Steller’s sea cow, this assembly will also allow future comparative studies of sirenians and other marine mammals.

Methods

Biological materials

The sample was collected from a wild adult female dugong captured as part of an ongoing research program in Moreton Bay, Queensland, Australia (−27.15148032, 153.0415985) on 17 May 2022. A total volume of 16 ml of whole blood in EDTA was collected nonlethally and immediately flash frozen on 17 May 2022. A total volume of 16 ml of whole blood in EDTA was collected nonlethally and immediately flash frozen in liquid nitrogen and stored at −80 °C until genomic DNA extraction. Samples were collected under Scientific Purposes Permit # WA0019236, Moreton Bay Marine Park permit # MPP18-001119, and UQ Animal Ethics permit # 2021/AE00821.

Nucleic acid extraction

We isolated high molecular weight (HMW) genomic DNA (>40 kbp) using a Circulomics Nanobind CBB kit (Pacific Biosciences—PacBio, Cat. #102-207-600). Prior to library preparation, the genomic DNA was pretreated for damage using the NEBNext FFPE DNA Repair Mix (New England Biolabs, Massachusetts), according to the manufacturer’s instructions.

PacBio HiFi library preparation and sequencing

Two HiFi SMRTbell libraries were constructed using the SMRTbell Express Template Prep Kit v2.0 (PacBio, Cat. #100-938-900) according to the manufacturer’s instructions. HMW gDNA was sheared to a target DNA size distribution between 15 and 20 kbp. The sheared gDNA was concentrated using 0.45× of AMPure PB beads (PacBio, Cat. #100-265-900) for the removal of single-strand overhangs at 37 °C for 15 min, followed by further enzymatic steps of DNA damage repair at 37 °C for 30 min, end repair and A-tailing at 20 °C for 10 min and 65 °C for 30 min, ligation of overhang adapter v3 at 20 °C for 60 min and 65 °C for 10 min to inactivate the ligase, then nuclease treated at 37 °C for 1 h. The SMRTbell library was purified and concentrated with 0.45× Ampure PB beads (PacBio, Cat. #100-265-900) for size selection using the BluePippin/PippinHT system (Sage Science, Massachusetts; Cat. #BLF7510/HPE7510) to collect fragments greater than 7 to 9 kbp. The 15 kbp average HiFi SMRTbell libraries were sequenced at the Australian Genome Research Facility in the University of Queensland using 3 8M SMRT cells, Sequel II sequencing chemistry 2.0, and 30-h movies each on a PacBio Sequel II sequencer.

Omni-C library preparation and sequencing

The Omni-C library was prepared from 3 ml of frozen blood using Dovetail Omni-C Kit (Dovetail Genomics, California) according to the manufacturer’s Mammalian protocol v1.4 with minor modifications. In brief, cells were isolated from thawed blood and chromatin fixed in place in the nucleus. Fixed chromatin was digested with DNase I then extracted and digestion profiles were assessed using TapeStation D5000 screen tapes (Agilent Technologies, California). Chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter-containing ends. After proximity ligation, crosslinks were reversed and the DNA purified from proteins. Purified DNA was treated to remove biotin that was not internal to ligated fragments. An NGS library was generated using an NEB Ultra II DNA Library Prep kit (New England Biolabs, Massachusetts) with an Illumina-compatible y-adapter. Biotin-containing fragments were then captured using streptavidin beads. The post-capture product was split into two replicates prior to PCR enrichment to preserve library complexity with each replicate receiving unique dual indices. The libraries were then sequenced at the Ramaciotti Center for Genomics at the University of New South Wales (Sydney, Australia) on an Illumina NextSeq 500 platform to generate approximately 100 million 2 × 150 bp read pairs per Gbp genome size.

Nuclear genome assembly

We assembled the dugong genome following the Vertebrate Genomes Project (VGP) v2.0 Galaxy assembly pipeline (Table 1, see Data availability statement for link to all assembly scripts) (Rhee et al. 2021; Lariiviére et al. 2023). In particular, we removed remnant adapter sequences from the PacBio HiFi dataset using cutadapt (Martin 2011) and used them to generate the initial phased diploid contigs using HiFiasm in HiC mode, with Omni-C used to phase the haplotypes (Cheng et al. 2021). We scaffolded both contig haplotypes using the Omni-C data with YaHS (Zhou et al. 2023). We generated Omni-C contact maps for both assemblies by aligning the Omni-C data against the corresponding assembly with
We assembled the mitochondrial genome of the dugong from the PacBio HiFi reads using the reference-guided pipeline MitoHiFi (https://github.com/marcelauliano/MitoHiFi) (Uliano-Silva et al. 2023). A previously assembled dugong mitogenome (NCBI:AY075116.1) was used as the starting reference sequence. After completion of the nuclear genome, we searched for matches of the resulting mitochondrial assembly sequence in the nuclear genome assembly using BLAST+ (Camacho et al. 2009) and filtered out contigs and scaffolds from the nuclear genome with a percentage of sequence identity >99% and size smaller than the mitochondrial assembly sequence. We annotated the resulting mitochondrial assembly using GeSeq (Tillich et al. 2017), implementing the tRNAscan-SE v.2.0.7 3rd party tRNA annotator with the vertebrate mitochondrial tRNA database for tRNA annotation.

**Genome size estimation and quality assessment**

We generated k-mer counts from the PacBio HiFi reads using meryl (https://github.com/marbl/meryl). We then applied GenomeScope 2.0 (Ranallo-Benavidez et al. 2020) to the k-mer database to estimate genome features including
Diversity and demographic history
We used ROHan (Renaud et al. 2019) on the filtered and aligned Omni-C data to refine estimates of genome-wide heterozygosity and identify runs of homozygosity (ROH), indicative of inbreeding. We applied the pairwise sequentially Markovian coalescent (PSMC) (Li and Durbin 2011) approach to infer historical effective population size of dugongs over time. We generated a diploid consensus sequence using the mpileup function of SAMtools (v0.1.18; with “-C50” option), bcftools to call variants, and available scripts from PSMC package to convert file formats. We required that sequencing depth for each locus was above one-third of average coverage (“-d” option) and less than twice of average coverage (“-D” option), and that consensus base quality was above Q20. We ran PSMC using the recommended parameters (Table 1) and 100 rounds of bootstrapping. We scaled our estimates using the previously reported dugong generation time of 27 yr (McDonald 2005) and a mutation rate of 6.25e-9 mutations per nucleotide per generation, calculated using the divergence rate between modern dugongs and Steller’s sea cows (Le Duc et al. 2022).

Discussion
We present a draft genome assembly for the culturally important dugong, assembled using long reads, and chromosome-scale sequencing data. Genome assemblies are available on NCBI for two other Sirenians, the Florida subspecies of the West Indian manatee (Trichechus manatus) or West African (Trichechus senegalensis) manatee species, both of which are listed as Vulnerable by the IUCN. Our assembly is the most contiguous sirenian genome assembly to date, improving on previous assemblies—all assembled with short-read data—by at least an order of magnitude in contigs and scaffold N50s.

Initial estimates of genome-wide heterozygosity based on our new genome assembly are relatively high for a mammal of
conservation concern (Fig. 2A), probably reflecting the previously high abundance of dugongs prior to the last interglacial period (ca. 125,000 yr ago). While ROH indicate past inbreeding, we find no evidence in the genome of ongoing inbreeding among the Moreton Bay population of dugongs where this reference individual was sourced from. Future analyses of individuals from different populations may show whether these patterns of diversity are replicated in smaller and more isolated populations.

Our demographic inference analysis based on PSMC suggests that dugongs in Eastern Australia were variably

### Table 2. Sequencing and assembly statistics, and accession numbers.

<table>
<thead>
<tr>
<th>BioProjects and vouchers</th>
<th>VGP NCBI BioProject</th>
<th>PRJNA489243</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species NCBI BioProject</td>
<td>PRJNA970804</td>
<td></td>
</tr>
<tr>
<td>NCBI BioSample</td>
<td>SAMN33212336</td>
<td></td>
</tr>
<tr>
<td>NCBI Genome accessions</td>
<td>Haplotype 1</td>
<td>Haplotype 2</td>
</tr>
<tr>
<td>Assembly accession</td>
<td>GCA_030035585.1</td>
<td>GCA_030020955.1</td>
</tr>
<tr>
<td>Genome sequences</td>
<td>JASCZL000000000</td>
<td>JASCZM000000000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome sequence</th>
<th>PacBio HiFi reads</th>
<th>Run</th>
<th>3 PACBIO_SMRT (Sequel II) runs: 6.5 million reads, 102 Gbases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Omni-C Illumina reads</td>
<td>Run</td>
<td>2 ILLUMINA (Illumina NovaSeq 6000) runs: 457.5 million reads, 138.2Gb</td>
<td></td>
</tr>
<tr>
<td>Assembly identifier (quality code)</td>
<td>mDugDug1 1(8.8.P8.Q70.C99)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HiFi read coverageb</td>
<td></td>
<td>32.0X</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genome Assembly Quality Metrics</th>
<th>Haplotype 1</th>
<th>Haplotype 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of contigs</td>
<td>294</td>
<td>256</td>
</tr>
<tr>
<td>Contig N50 (bp)</td>
<td>57,632,671</td>
<td>57,883,746</td>
</tr>
<tr>
<td>Contig NG50 (bp)</td>
<td>57,632,671</td>
<td>57,883,746</td>
</tr>
<tr>
<td>Longest contig</td>
<td>162,184,114</td>
<td>209,448,431</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>198</td>
<td>167</td>
</tr>
<tr>
<td>Scaffold N50 (bp)</td>
<td>177,379,183</td>
<td>138,031,769</td>
</tr>
<tr>
<td>Scaffold NG50 (bp)</td>
<td>177,379,183</td>
<td>138,031,769</td>
</tr>
<tr>
<td>Largest scaffold</td>
<td>267,865,978</td>
<td>230,272,189</td>
</tr>
<tr>
<td>Size of final assembly (bp)</td>
<td>3,159,179,246</td>
<td>3,154,861,630</td>
</tr>
<tr>
<td>Phased block NG50 (bp)</td>
<td>57,632,671</td>
<td>57,883,746</td>
</tr>
<tr>
<td>Gaps per Gbp (# Gaps)</td>
<td>25 (79)</td>
<td>28 (88)</td>
</tr>
<tr>
<td>Indel QV (frameshift)</td>
<td>41.52</td>
<td>42.16</td>
</tr>
<tr>
<td>Base pair QV</td>
<td>70.4553</td>
<td>70.3254</td>
</tr>
<tr>
<td>K-mer completeness</td>
<td>97.9001</td>
<td>97.8847</td>
</tr>
<tr>
<td>BUSCO completeness (vertebrata), n = 3354</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vertebrata n = 3354</td>
<td>H1×</td>
<td>97.9% 95.9% 2.0% 1.0% 1.1%</td>
</tr>
<tr>
<td>Mammalia n = 9226</td>
<td>H1×</td>
<td>97.8% 95.7% 2.1% 1.1% 1.1%</td>
</tr>
<tr>
<td>Organelles</td>
<td>H1×</td>
<td>96.2% 95.3% 0.9% 0.8% 3.0%</td>
</tr>
<tr>
<td></td>
<td>H2×</td>
<td>96.1% 95.2% 0.9% 0.8% 3.1%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1 complete mitochondrial sequence (pending NCBI accession code)</td>
</tr>
</tbody>
</table>

aAssembly quality code x-y-P-Q-C derived notation, from (Rhie et al. 2021). x = \log_{10}[^{contig NG50}]; y = \log_{10}[scaffold NG50]; P = \log_{10}[phased block NG50]; Q = Phred base accuracy QV (Quality value); C = % genome represented by the first “n” scaffolds, following a karyotype of 2n = 48 inferred from ancestral taxa Trichechus manatus (Noronha et al. 2022).
bRead coverage and NGx statistics have been calculated based on the estimated genome size of 3.16 Gbp.
cComplete BUSCOs (C), Complete and single-copy BUSCOs (S), Complete and duplicated BUSCOs (D), Fragmented BUSCOs (F), Missing BUSCOs (M).
d(H1) Haplotype 1 and (H2) Haplotype 2 assembly values.
abundant from around 1 million yr ago (Ma) to 150 ka. This earlier estimate coincides with the mid-Pleistocene transition, during which longer and more intense glacial cycling began. However, more recent fluctuations in dugong abundance do not precisely track the approximately 100 ka glacial cycles that drove changes in global sea level (Yehudai et al. 2021). Dugong abundance declined steeply beginning at ~100 ka, probably due to population fragmentation (Blair et al. 2014) and habitat loss that occurred as sea levels fell after the last interglacial period and the shallow seagrass meadows in which they lived disappeared.

Our draft genome assembly promises to advance understanding of marine mammal evolution and diversification as well as provide crucial insights into dugong conservation and management. Sirenians are the most ancient lineage of marine mammals, having split from their most recent terrestrial ancestor ~63.9 Ma (Yuan et al. 2021). Future comparative genomic studies both within Sirenia and between sirenians and other marine mammal lineages will shed light on the genomic changes that allowed for these lineages to adapt to the marine environment. For example, a more contiguous dugong reference genome will improve reference-guided assembly of the extinct Steller’s sea cow, which was notable for both its large size and its adaptation to a subpolar kelp forest environment, unique among the typically warm water dwelling Sirenia. Future generation of genome data from other dugong populations, many of which are geographically isolated and/or live in quite different environments, will allow evolutionary analyses of adaptations unique to this lineage. The species’ large but discontinuous geographic range raises the possibility that some populations are genetically distinct and locally adapted. By identifying isolated populations and better-defining subpopulation units, future work will allow development of more targeted management strategies that can support the continued persistence of this unique marine mammal in changing global habitats.

**Supplementary material**

Supplementary material is available at *Journal of Heredity* online.

**Acknowledgments**

We acknowledge the traditional custodians of the land and water, the Quandamooka people, who care for the yangang (dugong) and the Sea Country where the reference individual was sampled. We pay our respects to their elders past and present and recognize their ongoing connection between culture and Country. Thanks also to E. McLennan for undertaking the genomic DNA extractions. The authors wish to acknowledge the services of the Australian Genome Research Facility and the Ramaciotti Centre for Genomics.

**Funding**

Sample extractions and QC were facilitated by the Australian Research Council grant to KB (CE200100012). Funding for
the dugong reference genome was provided by grants from The Nature Conservancy to BS and HHMI to EDJ and BS and from the German Research Foundation to DLD and TS (SCHO624/13-1).

Conflict of interest statement. None declared.

Data availability

Data generated for this study are available under NCBI BioProject PRJNA970804. Raw PacBio HiFi and Omni-C Illumina sequencing data for NCBI BioSample SAMN33212336 are available at https://genomark.s3.amazonaws.com/index.html?prefix=species/Dugong_dugon/mDugDug1/germic/data/, pending submission to the NCBI Short Read Archive (SRA). GenBank accessions for both primary and alternate assemblies are GCA_030035585.1 and GCA_030020955.1. The mitochondrial genome is available at https://genomark.s3.amazonaws.com/index.html?prefix=species/Dugong_dugon/mDugDug1/assemblies_MT_rockerfeller/ pending submission to GenBank. Assembly scripts and other data for the analyses presented can be found at the VGP galaxy project: https://galaxyproject.org/projects/vgp/

References


