Received: 1 June 2020 | Revised: 8 September 2020 | Accepted: 8 October 2020

DOI: 10.1111/1755-0998.13284

FROM THE COVER

MOLECULAR ECOLOGY RESOURCES WILEY

Reference genome and demographic history of the most endangered marine mammal, the vaquita

Phillip A. Morin[1] | Frederick I. Archer[1] | Catherine D. Avila[2] | Jennifer R. Balacco[3] | Yury V. Bukhman[4] | William Chow[5] | Olivier Fedrigo[3] | Giulio Formenti[3] | Julie A. Fronczek[2] | Arkarachai Fungtammasan[6] | Frances M. D. Gulland[7] | Bettina Haase[3] | Mads Peter Heide-Jorgensen[8] | Marlys L. Houck[2] | Kerstin Howe[5] | Ann C. Misuraca[2] | Jacquelyn Mountcastle[3] | Whitney Musser[9] | Sadye Paez[10] | Sarah Pelan[5] | Adam Phillippy[11] | Arang Rhie[11] | Jacqueline Robinson[12] | Lorenzo Rojas-Bracho[13] | Teri K. Rowles[14] | Oliver A. Ryder[2] | Cynthia R. Smith[9] | Sacha Stevenson[9] | Barbara L. Taylor[1] | Jonas Teilmann[15] | James Torrance[5] | Randall S. Wells[16] | Andrew J. Westgate[17] | Erich D. Jarvis[10,18]

[1]Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, La Jolla, CA, USA

[2]San Diego Zoo Institute for Conservation Research, Escondido, CA, USA

[3]Vertebrate Genome Laboratory, The Rockefeller University, New York, NY, USA

[4]Regenerative Biology, Morgridge Institute for Research, Madison, WI, USA

[5]Wellcome Sanger Institute, Cambridge, UK

[6]DNAnexus, Mountain View, CA, USA

[7]University of California, Davis, Davis, CA, USA

[8]Greenland Institute of Natural Resources, Copenhagen K, Denmark

[9]National Marine Mammal Foundation, San Diego, CA, USA

[10]Laboratory of Neurogenetics of Language, The Rockefeller University, New York, NY, USA

[11]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, MD, USA

[12]Institute for Human Genetics, University of California, San Francisco, CA, USA

[13]Comisión Nacional de Áreas Naturales Protegidas/SEMARNAT, Ensenada, Mexico

[14]Office of Protected Resources, National Marine Fisheries Service, NOAA, Silver Spring, MD, USA

[15]Marine Mammal Research, Department of Bioscience, Aarhus University, Roskilde, Denmark

[16]Chicago Zoological Society's Sarasota Dolphin Research Program, c/o Mote Marine Laboratory, Sarasota, FL, USA

[17]University of North Carolina Wilmington, Wilmington, NC, USA

[18]Howard Hughes Medical Institute, Chevy Chase, MD, USA

Correspondence
Phillip A. Morin, Southwest Fisheries Science Center, National Marine Fisheries Service, NOAA, 8901 La Jolla Shores Dr., La Jolla, CA 92120, USA.
Email: phillip.morin@noaa.gov

Abstract

The vaquita is the most critically endangered marine mammal, with fewer than 19 remaining in the wild. First described in 1958, the vaquita has been in rapid decline for more than 20 years resulting from inadvertent deaths due to the increasing use of large-mesh gillnets. To understand the evolutionary and demographic history of the

vaquita, we used combined long-read sequencing and long-range scaffolding methods with long- and short-read RNA sequencing to generate a near error-free annotated reference genome assembly from cell lines derived from a female individual. The genome assembly consists of 99.92% of the assembled sequence contained in 21 nearly gapless chromosome-length autosome scaffolds and the X-chromosome scaffold, with a scaffold N50 of 115 Mb. Genome-wide heterozygosity is the lowest (0.01%) of any mammalian species analysed to date, but heterozygosity is evenly distributed across the chromosomes, consistent with long-term small population size at genetic equilibrium, rather than low diversity resulting from a recent population bottleneck or inbreeding. Historical demography of the vaquita indicates long-term population stability at less than 5,000 (*Ne*) for over 200,000 years. Together, these analyses indicate that the vaquita genome has had ample opportunity to purge highly deleterious alleles and potentially maintain diversity necessary for population health.

KEYWORDS

Conservation genomics, genome diversity, historical demography, Phocoena sinus, porpoise, Vertebrate Genomes Project

1 | INTRODUCTION

On the afternoon of 4 November 2017, an adult female vaquita porpoise (*Phocoena sinus*), the smallest and rarest cetacean in the world, was captured in a massive effort to save the species by bringing into captivity as many as possible of the estimated maximum of 30 remaining individuals at the time (Thomas et al., 2017). This represented only the second live capture of a vaquita ever, the first of which, just a few weeks earlier as part of the same conservation effort, resulted in release of the animal after only hours when it showed signs of continuing stress. Despite the efforts of an international team of scientists and experts in porpoise capture and care, the second captured vaquita (V02F) suffered stress-induced cardiac failure and died approximately seven hours after initial capture (Rojas-Bracho et al., 2019). That death ended the effort by the Vaquita Conservation, Protection, and Recovery (VaquitaCPR) project to temporarily protect vaquita near their native habitat in the northern Gulf of California, near San Felipe, Mexico. However, the careful planning and presence of veterinarian experts in marine mammal stranding response allowed for an immediate necropsy that went through the night, with harvest and storage of ovaries and other tissues for delivery to facilities 260 miles north near San Diego, California for tissue culture and cryopreservation. By 8:00 p.m. the next day, within 24 hr of the animal's cardiac arrest, the tissues were delivered to the Institute for Conservation Research, San Diego Zoo Global, for the culture of cells from as many tissues as possible. After weeks of tissue culture, cells were harvested and banked for future research, and frozen samples sent to the Vertebrate Genome Laboratory at The Rockefeller University to extract ultra-high molecular weight DNA and RNA for genome sequencing, assembly and transcriptome annotation.

This extraordinary effort to extract as much information as possible from the VaquitaCPR project reflects the broad scientific value placed on biodiversity and conservation. Sequencing of reference genomes is increasingly recognized as an important contribution to identify, characterize and conserve biodiversity (Garner et al., 2016; Harrisson et al., 2014; He et al., 2016; Morin et al., 2020; Supple & Shapiro, 2018), especially for species that are naturally rare and difficult to study. Reference genomes provide primary data to understand evolutionary relationships (Arnason et al., 2018; Zhou et al., 2018), historical demography (Armstrong et al., 2019; Foote et al., 2016; Morin, Foote, Baker, et al., 2018; Robinson et al., 2016; Westbury et al., 2019), evolution of genes and traits (Autenrieth et al., 2018; Fan et al., 2019; Foote et al., 2015; Morin et al., 2020; Springer, Emerling, et al., 2016; Springer, Starrett, et al., 2016; Yim et al., 2014) and susceptibility to inbreeding and outbreeding depression (Chattopadhyay et al., 2019; Hedrick et al., 2019; Robinson et al., 2018; Tunstall et al., 2018). Genomic resources also provide the tools for broader studies of population structure, relatedness and potential for recovery (e.g., Garner et al., 2016; Morin, Foote, Hill, et al., 2018; Tunstall et al., 2018).

The vaquita was described for the first time in 1958 (Norris & McFarland, 1958) and has been characterized as a naturally rare endemic species, limited to shallow, turbid and highly productive habitat in the upper Gulf of California between Baja California and mainland Mexico (Rodriguez-Perez et al., 2018). The vaquita's closest relatives are the congeneric Burmeister's (*P. spinipinnis*) and spectacled (*P. dioptrica*) porpoises, which are found only in temperate and cold waters in the Southern Hemisphere, separated by at least 5,000 km of ocean and two million years of divergence (Ben Chehida et al., 2020; McGowen et al., 2009; Rosel et al., 1995). Similar to other porpoises, vaquitas become entangled and die in gillnets set for finfish and shrimp (Rojas-Bracho &

Reeves, 2013). The mortality rate was known to be unsustainable when studies on the bycatch rate (D'Agrosa et al., 2000) and life history (Hohn et al., 1996) were combined with the first abundance estimate of $N = 567$ individuals (95% CI: 177–1073) in 1997 (Jaramillo-Legorreta et al., 1999). The rate of decline has increased since approximately 2011 due to entanglement in illegal gillnets targeting totoaba (*Totoaba macdonaldi*), a large fish approximately the same size as the vaquita, captured for the black market trade of their swim bladders in China (Rojas-Bracho et al., 2019). The International Union for the Conservation of Nature (IUCN) has listed the vaquita as critically endangered since 1996, and the most recent estimates from 2018 indicate that fewer than 19 vaquita survive (Jaramillo-Legorreta et al., 2019). Initial genetics studies found no variation in mitochondrial DNA (mtDNA; Rosel & Rojas-Bracho, 1999) and low variation in the MHC DRB locus (Munguia-Vega et al., 2007). These authors have suggested that the low genetic diversity is due to long-term low effective population size (*Ne*) rather than to a recent bottleneck or the current rapid population decline (Munguia-Vega et al., 2007; Rojas-Bracho & Taylor, 1999; Taylor & Rojas-Bracho, 1999), but these data from few loci provide limited power to estimate timing or duration of demographic changes.

As part of the effort to prevent extinction of the vaquita and to further develop genomic resources to facilitate conservation and management planning for this and other endangered species, we used the Vertebrate Genomes Project (VGP) pipeline to generate a chromosomal-level, haplotype-phased reference vaquita genome assembly that exceeds the "platinum-quality" reference standards established by the VGP (Rhie, McCarthy, et al., 2020). The VGP standards are guidelines to ensure minimum error rates (QV40 or higher, or no more than 1 nucleotide error per 10,000 bp), highly contiguous and complete assemblies (contig N50 ≥ 1 Mb; chromosomal scaffold N50 ≥ 10 Mb), phasing of paternal and maternal haplotypes to reduce false gene duplication errors and manual curation to reduce errors and improve genome assembly quality. Based on the reference-quality assembly, we analysed genomic diversity and historical demography to infer the cause of current low genomic diversity and whether genetic factors should be considered to be of concern for recovery if the immediate reason for decline, incidental bycatch in gillnets, can be halted in time to prevent extinction.

2 | MATERIALS AND METHODS

2.1 | Genome data generation

Skin, mesovarium, kidney, trachea, and liver tissues were obtained during necropsy of the adult female vaquita that died during an attempt to begin ex-situ protection from illegal fishing operations (Rojas-Bracho et al., 2019). Cells were harvested and cultured at the Institute for Conservation Research, San Diego Zoo Global (Frozen Zoo®). From these cells, we generated a reference quality genome using the VGP pipeline 1.5 (Rhie, McCarthy, et al., 2020). In particular, we collected four genomic data types: Pacific Biosciences (Menlo Park, CA, USA) continuous long reads (CLR), 10X Genomics (Pleasanton, CA, USA) linked-reads, Bionano Genomics, Inc. (San Diego, CA, USA) DLS optical maps, and Arima Genomics, Inc. (San Diego, CA, USA) v1 Hi-C data. From one tube containing ~4 million cells in 1× PBS buffer (137 mM NaCl, 10 mM phosphate, 2.7 mM KCl; pH 7.4) with 10% DMSO and 10% Glycerol, ultra-high molecular weight DNA (uHMW DNA) was extracted using the agarose plug Bionano Genomics protocol for Cell Culture DNA Isolation (Bionano Genomics, document No. 30026F). uHMW DNA quality was assessed by a Pulsed Field Gel assay and quantified with a Qubit 2 Fluorometer. From these extractions, 10 μg of uHMW DNA was sheared using a 26G blunt end needle (PacBio protocol PN 101-181-000 Version 05). A large-insert PacBio library was prepared using the Pacific Biosciences Express Template Prep Kit version 1.0 (PN 101-357-000) following the manufacturer protocol. The library was then size selected (>20 kb) using the Sage Science BluePippin Size-Selection System and sequenced on 30 PacBio 1M version 3 SMRT cells on the Sequel I instrument with the sequencing kit 3.0 (PN 101-597-800) and 10 hr movie. We used the same unfragmented DNA to generate a 10X Genomics Chromium linked-reads library (Genome Library Kit & Gel Bead Kit version 2, PN 120258, Genome HT Library Kit & Gel Bead Kit version 2, PN 120261, Genome Chip Kit version 2, PN 120257, i7 Multiplex Kit, PN 120262). We sequenced this 10X Genomics library on an Illumina Novaseq S4 150 bp PE lane.

An aliquot of the same DNA was labelled for Bionano Genomics optical mapping using the Bionano Prep Direct Label and Stain (DLS) Protocol (document No. 30206E) and run on one Saphyr instrument chip flowcell. Hi-C reactions were performed by Arima Genomics according to the protocols described in the Arima-HiC kit (PN A510008). After the Arima-HiC protocol, Illumina-compatible sequencing libraries were prepared by first shearing purified Arima-HiC proximally-ligated DNA and then size-selecting DNA fragments from ~200–600 bp using SPRI beads. The size-selected fragments were then enriched for biotin and converted into Illumina-compatible sequencing libraries using the KAPA Hyper Prep kit (PN KK8504). After adapter ligation, DNA was PCR amplified and purified using SPRI beads. The purified DNA underwent standard QC (qPCR and Bioanalyser [Agilent]) and was sequenced on the Illumina HiSeq X to ~60X coverage following the manufacturer's protocols.

2.2 | Transcriptome data generation

Total RNA extraction and purification was conducted with QIAGEN RNeasy kit (PN 74104). The quality and quantity of all RNAs were measured using a Fragment Analyzer (Agilent Technologies, Santa Clara, CA) and a Qubit 2.0 (Invitrogen). PacBio Iso-Seq libraries were prepared according to the "Procedure & Checklist - Iso-Seq Template Preparation for Sequel Systems" (PN 101-763-800 Version 01). Briefly, cDNA was reverse transcribed using the NEBNext Single Cell/Low Input cDNA Synthesis & Amplification Module (NEB E6421S) from

238 ng total RNA. Amplified cDNA was cleaned with 86 μl ProNex beads. The PacBio Iso-seq library was sequenced on one PacBio 8M (PN 101-389-001) SMRT Cell on the Sequel II instrument with sequencing kit 1.0 (PN 101-746-800) using the Sequel II Binding Kit 1.0 (PN 101-726-700) and 30 hr movie with two hours pre-extension.

The same RNA was used for mRNA-seq. The RNA-Seq library was prepared with 100 ng total RNA using the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, PN E7490S) followed by NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (PN E7760S). The library was then amplified over 14 cycles. Library quantification and qualification were performed with the Invitrogen Qubit dsDNA HS Assay Kit (PN Q32854). Libraries were sequenced on the Illumina NextSeq 500 in 150PE mid-output mode (Rockefeller Genomics Center). Data quality control was done using fastQC (version 0.11.5; https://qubeshub.org/resources/fastqc).

2.3 | Genome assembly and annotation

We assembled the vaquita genome using the VGP 1.5 pipeline on the DNAnexus cloud computing system (https://platform.dnanexus.com/). Briefly, this pipeline is composed of an assembly step, scaffolding step and final polishing step. First, we assembled raw PacBio data with Falcon 2.0.0/Falcon-unzip 1.1.0 (Chin et al., 2016). Then, we polished the primary and alternate phased haplotype contigs using the same PacBio reads with arrow (PacBio smrtanalysis 6.0.0.47841). Prior to scaffolding, we detected and reassigned haplotype duplicated contigs in the primary contig set using purge_haplotig 1.0.4 (Roach et al., 2018) and we also extracted the mitochondrial reads to assemble the mitochondral sequence (Formenti et al., 2020). From this step, we only scaffolded the primary haplotype contigs using 10X Genomics data with scaff10x 4.1 (https://github.com/wtsi-hpag/Scaff10X), Bionano CMAP with Bionano Hybrid Solve 3.3_10252018 (Bionano Genomics) and Hi-C data with Salsa 2.2 (Ghurye et al., 2017). Finally, the resulting primary scaffolds and alternate contigs were processed together through three polishing rounds: one additional round of arrow polishing and two rounds of polishing using 10X Illumina data mapped with Long Ranger version 2.2.2 (https://github.com/10XGenomics/longranger) and base calling with FreeBayes 1.2.0 (Garrison & Marth, 2012). Primary scaffolds and alternate contigs were contamination checked and curated manually (Howe et al., 2020) using gEVAL (Chow et al., 2016). For the primary assembly, this resulted in a further reduction of scaffold numbers by 11% and an increase of the scaffold N50 by 12% to 115 Mb. The scaffolds were aligned to the blue whale genome (NCBI accession GCA_009873245.2) using the Nucmer package in MUMmer version 3.0 (Kurtz et al., 2004) to assign scaffolds to chromosomes based on synteny. The primary and associated alternate assemblies were submitted to NCBI (accession GCA_008692025.1), and annotation was performed through their standard pipeline incorporating our RNA-seq and Iso-seq data (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/). The primary assembly was screened for repetitive elements using RepeatMasker version 4.0.5 (Smit et al., 2013–2015) and the RepeatMasker combined database Dfam_Consensus-20181026. Base accuracy (QV) was measured using k = 21 with Merqury (Rhie et al., 2020). Gene content of the primary scaffolds was assessed using BUSCO version 3.1.0 (Waterhouse et al., 2017) searches of the Laurasiatheria and mammalian gene set databases.

2.4 | Historical demography

To conduct analysis of historical demography using pairwise sequentially Markovian coalescent (PSMC; Li & Durbin, 2011), we first aligned 10X Genomics paired-end reads to the primary haplotype assembly (Armstrong et al., 2019) to create a diploid nuclear genome pileup. The reads were trimmed with the BBduk function of BBTools (sourceforge.net/projects/bbmap/), removing the first 22 nucleotides of the R1 reads introduced during the Chromium library preparation (https://support.10xgenomics.com/genome-exome/library-prep/doc/technical-note-assay-scheme-and-configuration-of-chromium-genome-v2-libraries) and trimming all reads for average quality (q ≥ 20), 3' ends trimmed to q ≥ 15 and minimum length (≥40 nucleotides). Unpaired reads were removed from the trimmed fastq files using the BBTools repair.sh function. Trimmed reads were aligned to the vaquita mitogenome (accession CM018178.1) using BWA mem (Li & Durbin, 2009), and the unmapped reads exported as reads representing only the nuclear genome. Nuclear reads were aligned to the primary haplotype assembly (accession GCA_008692025.1), and duplicate reads removed using Picard-Tools (http://broadinstitute.github.io/picard/). The resulting genome alignments from four 10X Genomics libraries were assessed for average depth of coverage using ANGSD (Korneliussen et al., 2014), and combined for 47.8X average depth of coverage. From this coverage pile-up, the diploid consensus genome was extracted (Li & Durbin, 2011) and used as input for PSMC with generation time of 11.9 years based on the estimated generation time of harbor porpoise (Taylor et al., 2007), and an autosomal mutation rate (μA) of 1.08×10^{-8} substitutions per nucleotide per generation based on a previously determined rate for odontocetes (9.1E-10 subst/site/yr; Dornburg et al., 2012). PSMC atomic time intervals were combined as suggested by the authors (https://github.com/lh3/psmc) such that after 20 rounds of iterations, at least ~10 recombinations are inferred to occur in the intervals each parameter spans: $p = (8 + 23 * 2 + 9 + 1)$. The remaining parameters were left as the default values used for humans (Li & Durbin, 2011), and we performed 100 bootstrap resamplings on all PSMC analyses to assess variance of the model. We also conducted PSMC after masking of repeat regions to control for potential bias due to variation in collapsed repeats (Patil et al., 2020), but because the removal of repeat regions resulted in significantly less data, the PSMC parameters could not be optimized for sufficient resolution in older time periods, resulting in higher variance for those intervals. The effects of collapsed repeats (expected to be less pronounced in a very complete genome like this one, where most repeats have been resolved) are expected to only change the *Ne* estimates in the most recent and oldest time periods (Patil et al., 2020). Our repeat-masked plot looks nearly identical to the full-genome plot in the most recent and middle time periods, and similar (but higher variance) in the oldest time periods (Figure S1).

2.5 | Genome-wide heterozygosity

The distribution of heterozygosity across the genome was determined using previously described analysis pipelines (Robinson et al., 2019). Briefly, we used HaplotypeCaller in the Genome Analysis Toolkit (GATK; McKenna et al., 2010) to call genotypes from the short-read pile-up (above), filtering out sites with <1/3X or >2X the average depth of coverage. Heterozygosity was calculated as the number of heterozygous sites divided by the total number of called genotypes in nonoverlapping 1 Mb windows across each scaffold.

2.6 | Modelling demographic effects on heterozygosity

A coalescent simulation was constructed to estimate recent effective population size (rN_e), historical effective population size (hN_e) and time since a bottleneck (b) in which the population reduced in size from hN_e to rN_e. The analysis computed the likelihood of the empirical distribution of the number of heterozygous sites per kb (H_{kb}) observed in 2,244 1 Mb windows in the vaquita genome (from above) given similar distributions drawn from an equivalent genome arising from random draws of each of these parameters, which were sampled as:

$$rN_e \sim \text{Uniform}(0, 3) \times 10^4$$
$$hN_e \sim \text{Uniform}(0, 9) \times 10^4$$
$$b \sim 10^{\text{Uniform}(0, 7)}$$

We initially drew 50,000 random values from these distributions. We then randomly selected 20,000 of these values where average growth rates ($(rN_e/ hN_e)/ b$) were less than 1.06, as values above this were considered to be biologically improbable (Taylor et al., 2019).

For each of the 20,000 scenarios, we generated one million independent SNPs for a single individual with a mutation rate of 1.08×10^{-8} substitutions/site/generation and a generation time of 11.9 years. To capture variability in the coalescent, we ran 4,488

TABLE 1 Vaquita genome assembly metrics. Genome size is the kmer estimate based on GenomeScope (version 1.0) analysis of the 10X Genomics data with k = 31. The BUSCO score is for complete genes identified from the mammalian single-copy conserved gene data set

Genome quality metric	
Contig N50	20.22 Mb
Scaffold N50 (max size)	115.47 Mb (185.85 Mb)
No. scaffolds (primary haplotype)	64
Base quality (QV)	40.9
Genes identified (BUSCO)	91.6%
Assembly size (ungapped)	2,363,494,880 bp
Assembly size (total)	2,371,540,524 bp
Genome size	2,667,451,016 bp

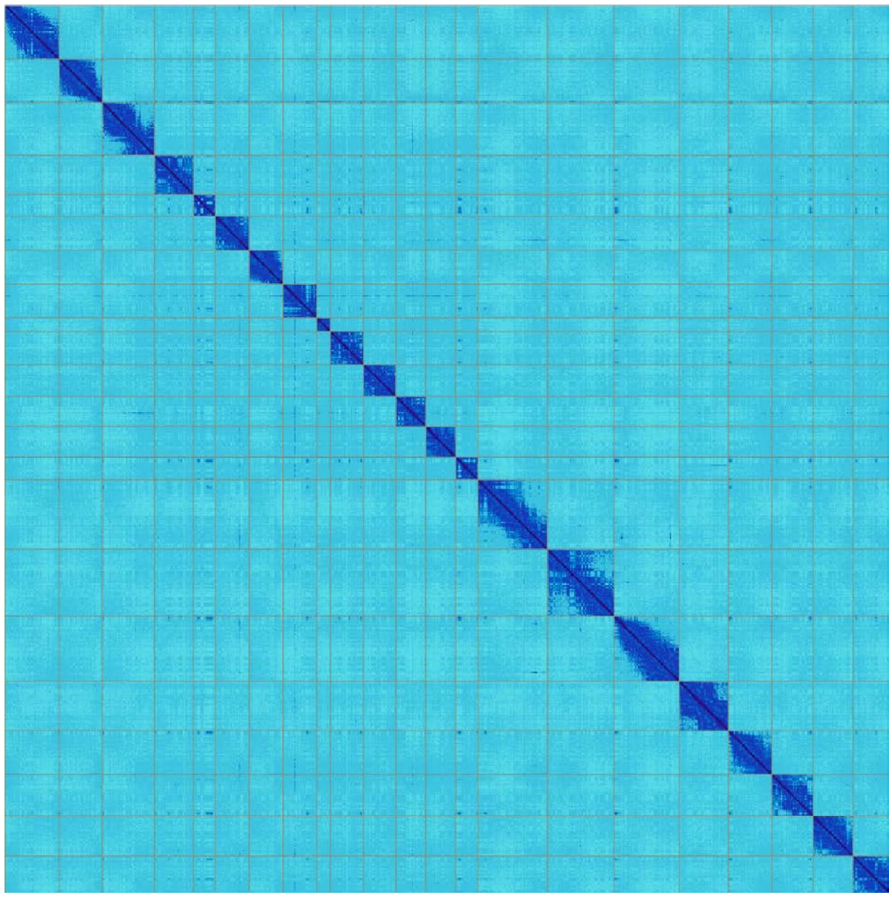

FIGURE 1 HiC heat-map of genomic interactions. Interactions between two locations are depicted by a dark blue pixel. Grey lines depict scaffold boundaries for the 22 chromosome-length scaffolds. Different scaffolds should not share any interactions (pixels off diagonal outside the scaffold boundaries), while patterns within a scaffold show chromosome-substructure

TABLE 2 Estimated genome sequence average depth of raw data coverage (before adapter and quality trimming) for sequencing and mapping technologies based on an estimated genome size of 2.7 Gb

Data type	Raw data (bp)	Coverage
10x Genomics	200,218,960,380	74X
Arima Genomics HiC	255,724,383,000	94X
Bionano Genomics	480,155,600,000	178X
PacBio SubReads	325,960,000,000	121X

replicates of each scenario, which was twice the number of ~1 Mb windows in the empirical vaquita genome. This ensured that we could produce enough random sets of 2,244 1 Mb windows from which to compute the scenario likelihoods as described below. The simulations were run with fastsimcoal version 2.6.2 (Excoffier et al., 2013) through the R package strataG (version 4.9.05).

For each of the 4,488 replicates of one million SNPs in a scenario, we calculated the number of heterozygous SNPs per KB (H'_{kb}). We then drew a random 2,244 values of H'_{kb} without replacement to represent one simulated genome for this scenario. We fit a gamma distribution to these values, which was used to compute the negative sum of log-likelihoods (-logL) of the empirical H_{kb} from the vaquita genome. For each scenario, we repeated this random draw of 2,244 values of H'_{kb} and computation of −logL 100 times and recorded the mean and standard deviation of -logL. Likelihoods were plotted as heatmaps of the LOESS smoothed fit of −logL across pairs of simulation parameters. LOESS models were fit to each pair of parameters separately, and the surfaces represent the predicted −logL of 100,000 (10,000 × 10,000) evenly spaced points across each plot.

3 | RESULTS

3.1 | A highly contiguous assembly of the vaquita genome

We assembled a 2.37 Gb genome (Table 1) in only 64 scaffolds, of which 21 represented arm-to-arm autosomes, named according to synteny with the blue whale (*Balaenoptera musculus*) and the X chromosome, in agreement with the 22-chromosome karyotype. The remaining 42 unplaced scaffolds consisted of only 0.198 Gb combined (0.08% of the total length), meaning that 99.92% of the assembled sequence has been assigned to chromosomes. Consistent with this mostly complete assembly, the N50 contig value was 20.22 Mb (273 contigs), N50 scaffold was 115.47 Mb, and base call accuracy was QV40.88 (0.82 errors per 10,000 bp). There were only 208 gaps, of which the annotated chromosomes had 3–17 gaps each. The Hi-C heat-map showing genomic interactions (Figure 1) indicates strong agreement between the close interactions and chromosome-length scaffolds. The alternate haplotype contigs are made up of 1 Gb of the genome, indicating low heterozygosity. Depth of coverage for each data type are presented in Table 2. Assemblies of both primary and alternate haplotypes have been deposited at DDBJ/ENA/GenBank under the accessions VOSU00000000 (principle haplotype) and VOSV00000000 (alternate haplotype) in BioProjects PRJNA557831 and PRJNA557832, respectively.

BUSCO analysis showed 89.9% and 91.6% gene content identification from the primary haplotype when compared to the Laurasiatheria and mammalian data sets, respectively, with only 1.0% and 1.1% of the complete genes duplicated, respectively, and 4.3% and 4.6% fragmented (Table S1). Genome annotation identified 26,497 genes and pseudogenes, 19,069 of which are protein coding (Table S2). The cumulative number of genes with alignment to the UniProtKB/Swiss-Prot curated proteins was 18,748 (89%) at ≥90% coverage of the target protein. This coverage was 5%–48% higher than the number of genes aligned from other annotated cetacean genomes (Table S2). Similar to other cetacean genomes (e.g., Fan et al., 2019; Keane et al., 2015; Tollis et al., 2019), the vaquita genome consisted of about 46% repeats (Table 3) based on RepeatMasker.

3.2 | Low heterozygosity of the vaquita genome

Genome-wide heterozygosity was 0.0105% overall (0.0097% for repeat-masked genome), with even distribution of heterozygosity across the genome (Figure 2a). Heterozygosity per 1 Mb window ranged from 0 to 1.2 sites/kb, but only two (noncontiguous) windows out of 2,247 had no heterozygotes, and the standard deviation of heterozygosity across the windows was very low ($SD = 0.0000767$). None of the 1 Mb windows had heterozygosity of >1.3 sites/kb, and 94% of the windows had heterozygosity of <0.2 sites/kb (Figure 2b). Thus, In comparison to other mammals, the vaquita genome exhibits the lowest heterozygosity yet detected in an outbreeding mammalian species (Figure 3, Table S3). The only lower heterozygosity was for one of the six subspecies of Island fox (*Urocyon littoralis*), the San Nicolas Island fox (U. l. *dickeyi*), an endemic island subspecies found only on a 58 km^2 island approximately 100 km off the coast of California, with an estimated population size of about 500 individuals (Robinson et al., 2016). However, unlike the vaquita, heterozygosity is not evenly distributed

TABLE 3 Repetitive content of the vaquita genome (total assembly length 2.372 Gb) as determined by RepeatMasker

Repeat type	Length (bp)	% of Genome
SINEs	189,109,608	7.97%
LINEs	653,546,597	27.56%
LTR	134,757,334	5.68%
DNA transposons	76,591,695	3.23%
Unclassified	1,047,864	0.04%
Satellites	1,588,863	0.07%
Simple repeats	23,753,228	1%
Low complexity	4,527,734	0.19%
Total repeats:	1,085,270,145	45.76%

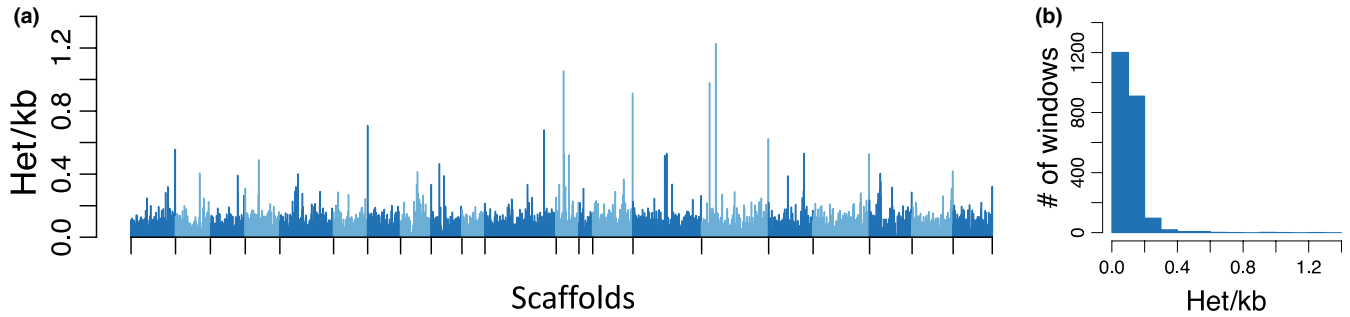

FIGURE 2 Distributions of heterozygosity across the vaquita genome. (a) Bar plot shows per-site heterozygosity in nonoverlapping 1-Mb windows across 21 autosomal scaffolds >10 Mb in length. Scaffolds are shown in alternating shades. Note that some of the highes peaks of heterozygosity are at the chromosme (scaffold) ends. (b) Histogram of the count of per-window heterozygosity levels

across the genome in the San Nicolas Island fox and other small inbred populations of canids, due to the effects of recent inbreeding in addition to long-term small population sizes (Robinson et al., 2019).

3.3 | Vaquita population size over time

This low, relatively even heterozygosity across the vaquita genome could be indicative of a long-term small, outbred population (Robinson et al., 2019; Westbury et al., 2019). To test this hypothesis, we performed PSMC analysis of historical demography. The results indicate that the vaquita effective population size has been small, ranging from about 1,400 to 3,200 for most of the last ~300,000 years (Figure 4a, full genome; Figure S1, repeat-masked genome). This finding corroborates previous conclusions based on single-locus analyses (Munguia-Vega et al., 2007; Taylor & Rojas-Bracho, 1999) but extends the duration of persistence of the species at low *Ne* to the mid Pleistocene, prior to the penultimate glacial period, the Saalian, which lasted from approximately 300,000 to 130,000 years ago.

To further test the long-term low population hypothesis, we performed coalescent simulations of genome-wide heterozygosity across a range of demographic scenarios. The range of negative log-likelihood (-logL) values in the 20,000 scenarios was −3238–84,523. Given the long tail of this distribution, which represented portions of the parameter space with low likelihood (high values of −logL; Figure S2), we randomly selected 10,000 scenarios with −logL < 6,000 to examine. The heatmap likelihood plots across pairs of simulation parameters showed that the empirical distribution of heterozygosity in the 2,244 1 Mb windows across the genome is most likely the result of a population with a recent effective population size (rN_e) of less than 2,500 and that the population most likely went through a bottleneck approximately one million years ago (Ma; Figure 4b). There was more uncertainty in the likelihood surface for historical effective population size (hN_e) prior to this bottleneck. However, for bottlenecks less than 100,000 years ago, *hNe* was estimated to be low and close to *rNe*, suggesting that population size was relatively low and constant during this period (Figure 4c). If the bottleneck is presumed to occur around 1 MYA, as indicated in Figure 4b, historical N_e was likely to have been approximately between 20,000 and 50,000 (Figure 4c and d).

4 | DISCUSSION

We have assembled the most complete cetacean genome to date, as measured by the low number of scaffolds, small number of gaps per chromosome scaffold, high percentage of scaffolds assigned to 22 chromosomes, cumulative number of genes with an alignment to the UniProtKB/Swiss-Prot curated proteins and small amount of missing data. Identification of gene content was also in the expected range for a high-quality mammalian genome at 90.5% of complete single-copy genes from the BUSCO mammalian gene set, with a low level of false duplicates and low levels of fragmented genes.

The PSMC analysis indicates that the vaquita population declined during the late Pleistocene, most likely due to climate change and the associated habitat changes in the eastern North Pacific coastal regions of North and Central America, and that it remained small over the last approximately 300,000 years. PSMC results can be affected by population structure, inbreeding, changes in connectivity among populations and stochastic variation in coalescent events when diversity is low (Beichman et al., 2017; Li & Durbin, 2011; Mazet et al., 2016; Orozco-terWengel, 2016). The coalescent results are consistent with the PSMC-inferred historical demography being the most likely cause of current heterozygosity levels rather than a recent severe bottleneck or inbreeding. Importantly, the duration of the small population size indicates that the observed level of heterozygosity is the result of a population at genetic equilibrium, where mutations are balanced by drift and selection, and that highly deleterious mutations are likely to have been purged from the population (Day et al., 2003; Robinson et al., 2018; Westbury et al., 2018, 2019).

Examples of species with low diversity but long-term viability and potential for adaptability are becoming more common (Foote et al., 2019; Robinson et al., 2018; Westbury et al., 2018, 2019; Xue et al., 2015). Among odontocetes (toothed whales, dolphins and porpoises), in particular, there are examples of species with nearly as low diversity as the vaquita that exhibit strong evidence of the influence

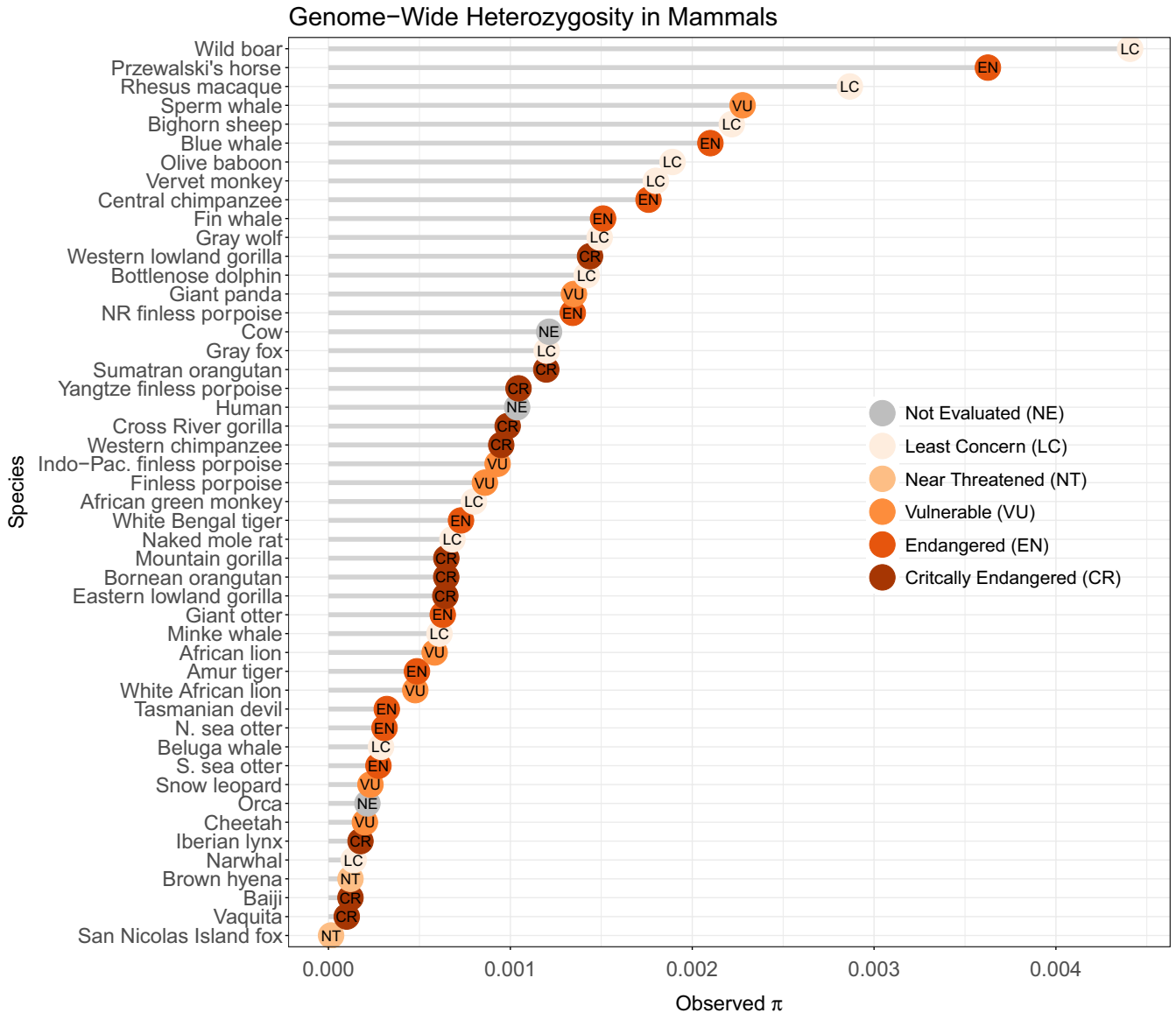

FIGURE 3 Comparison of genome-wide heterozygosity (π) among mammals. Values are drawn from the literature, based on Robinson et al. (2016), plus the vaquita and blue whale. Dots are coloured by the endangered status according to the Red List for Threatened Species, International Union for Conservation of Nature (IUCN). Although the Baiji, or Yangtze River dolphin, is listed as critically endangered, it is believed to have been extinct since at least 2006 (Turvey et al., 2007). See Table S4 for heterozygosity information

of demographic factors influencing genome-wide diversity over tens to hundreds of thousands of years of diversification and adaptation (Foote et al., 2016, 2019; Van Cise et al., 2019; Westbury et al., 2019). In several of these cases where it has been examined, genome-wide heterozygosity patterns do not indicate that low diversity was caused by rapid bottlenecks or inbreeding and is not associated with being endangered (Figure 3); instead, these patterns indicate that low diversity has been present for extended periods while species persist and diversify (e.g., narwhal (Westbury et al., 2019), orca (Foote et al., 2019)). These examples and others (Robinson et al., 2016, 2018; Westbury et al., 2018) indicate that, contrary to the paradigm of an "extinction vortex" (Gilpin & Soulé, 1986) that may doom species with low diversity, some species have persisted with low genomic diversity and small population size. Long-term small population size enables the purging of recessive deleterious alleles, thereby reducing the risk of inbreeding depression, perhaps allowing for continued future persistence with relatively small population sizes and an increased tolerance to the genetic consequences of bottlenecks.

The vaquita's current habitat in the upper Gulf of California was probably diminished or absent due to low sea levels several times through the last 350,000 years (Siddall et al., 2003), with the lowest sea level occurring at the end of the Saalian complex and the LGM (Figure 4) followed by a rapid rise of 120–140 m (similar to the present level) during the Eemian warm period between 115,000 and 130,000 years ago and after the LGM (Figure 5). Over much of the last 100,000 years, sea level has been intermediate between

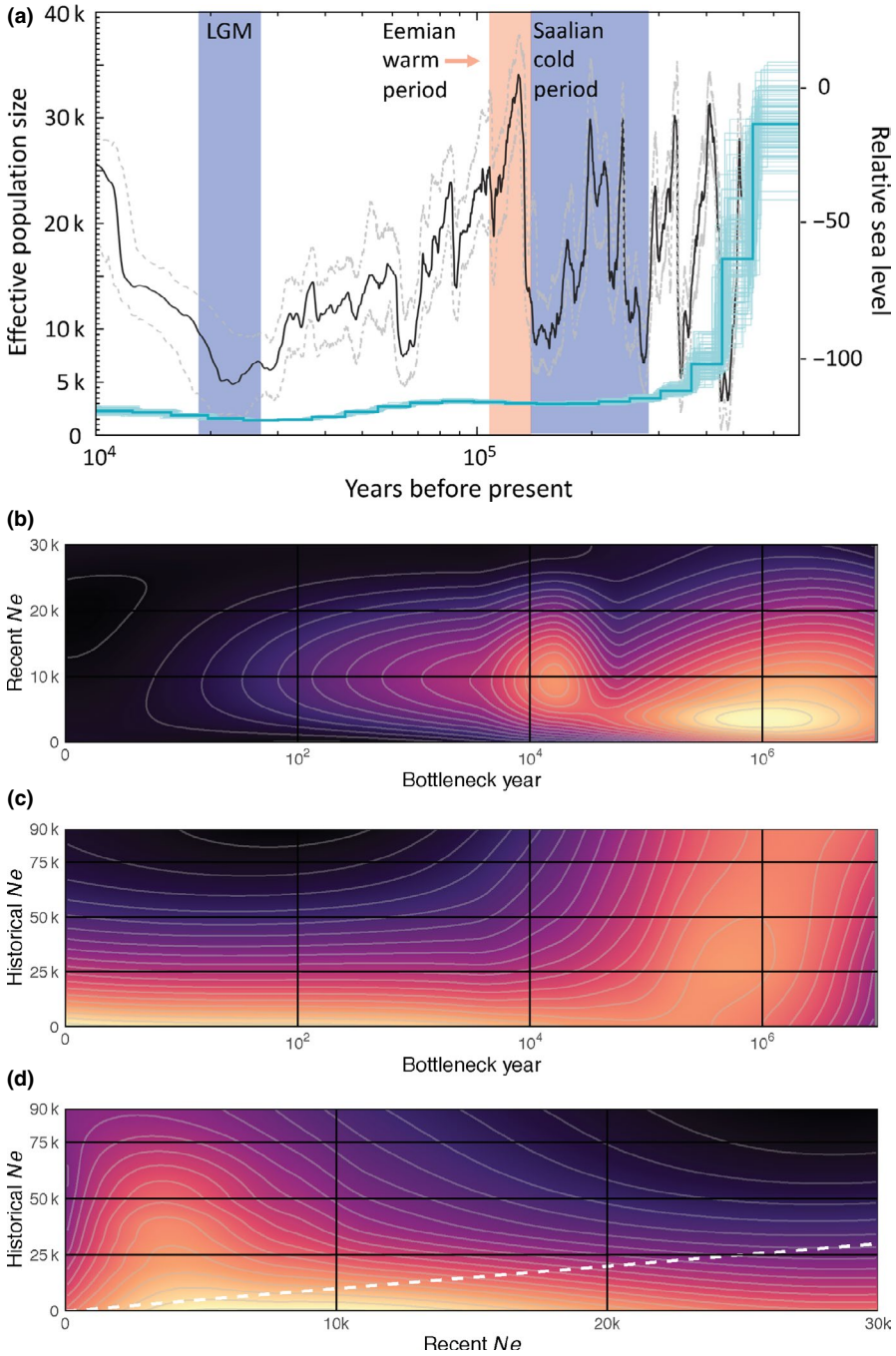

FIGURE 4 Changes in vaquita population size over time. (a) Changes in effective population size (*Ne*, *blue*, *left axis*) of the vaquita over time inferred from PSMC analysis of the nuclear genome. The darker blue line represents the median and lighter lines represent the 100 bootstrap replicates. The black line shows relative sea level (right axis, compared to present) with 95% confidence intervals (grey dashed lines) from Grant et al. (2014), and shading corresponds to cold and warm periods. (b–d) Heatmap of the distribution of the negative log-likelihood (-logL) of the empirical heterozygosity distribution across pairs of demographic parameters from the coalescent model, with higher likelihood combinations shown by lighter colour. The dashed white line in (d) represents a 1:1 slope, where current and historical population sizes would have been equal before and after the modelled change in population size

the high points (present and Eemian warm period) and lows (end of Saalian and the LGM; Rohling et al., 2017). There is no fossil record or other indication that vaquita have ever inhabited colder parts of the eastern North Pacific along the west coast of Baja California, Mexico, or further north off of California at the southern end of the current range of the congeneric harbour porpoise (*Phocoena phocoena*; Brownell, 1983). The closest relative of the vaquita, the Burmeister's porpoise or the ancestor of two sister species, Burmeister's and spectacled porpoise (Ben Chehida et al., 2020), are both found only in temperate and cold waters of the southern hemisphere. Based on the closer relationship to southern hemisphere species and on the similar timing of rapid climate warming and vaquita population decline, we hypothesize that climate change at the end of the Saalian ice age caused a northward shift of the species range, resulting in a remnant population being isolated in the Gulf of California, where it has persisted in the newly expanded and shallow, highly productive upper Gulf region.

The reference genome presented here has provided important insight into the demographic history of the critically endangered vaquita, reinforcing a previous hypothesis (Taylor & Rojas-Bracho, 1999) that the low genetic diversity of the vaquita is not due to a recent extreme bottleneck or current inbreeding. These results taken together with recent evidence of healthy looking vaquitas, often with robust calves (Taylor et al., 2019), suggest that population recovery may not be hindered because of genetic issues. Analysis of resequenced genomes from multiple individuals sampled over the previous few decades will shed light

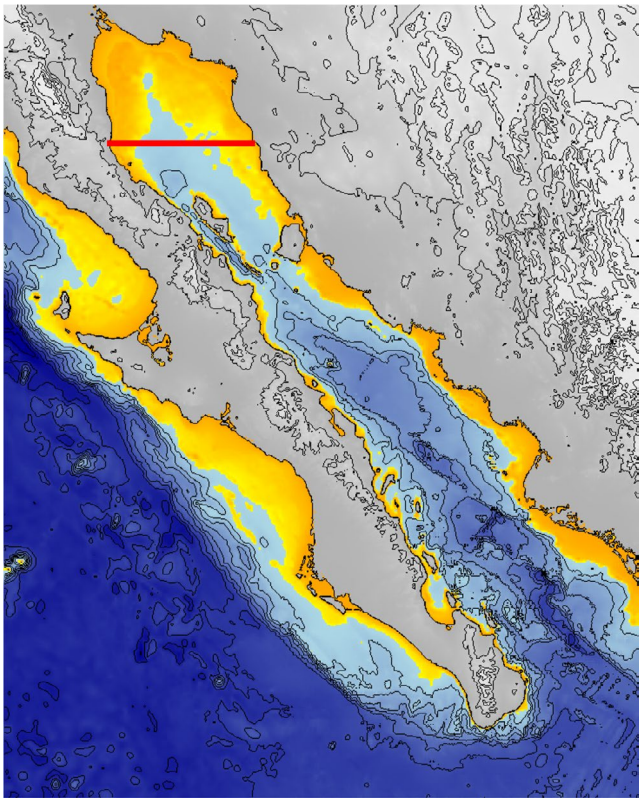

FIGURE 5 Bathymetric map of the Gulf of California showing 500 m isobath lines. Transition to yellow is at −140 m, indicating portions of the Gulf that were likely above sea level during the last two glacial maxima, ~22,000 and 140,000 years ago. The area north of the red line is the approximate historical range of the vaquita (Brownell, 1986)

on changes in inbreeding as the population has declined due to bycatch in gillnets, and whether deleterious mutations are likely to have been purged from the genome as a result of the long-term persistence at a small population size, as has been suggested for some other species and populations (e.g., Robinson et al., 2018; Westbury et al., 2018, 2019).

Finally, this genome assembly is the highest quality, most complete genome in the odontocete lineage that consists of all dolphins, porpoises and toothed whales. As such, it provides a genomic resource for better reference-guided assemblies and scaffolding of other cetacean genomes (Alonge et al., 2019; Lischer & Shimizu, 2017; Morin et al., 2020) and for comparative genomics, especially for variation in genome structure. We expect that the vaquita genome, along with expected assembly of reference genomes for other endangered species, will continue to contribute to both understanding and conservation of global biodiversity.

ACKNOWLEDGEMENTS

The planning and diligence of the VaquitaCPR team were critical in collection, preservation and rapid delivery of tissue samples for tissue culture that made this project possible. We are grateful to all people involved in obtaining and culturing the tissue samples. Kelly Robertson was instrumental in ensuring rapid import of the samples under CITES permit (permit No. 17US774233/9; Mexican export permit MX89760). Tissue samples are stored in the SWFSC Marine Mammal and Sea Turtle Research (MMASTR) collection under MMPA permit 19091. We thank Annabel Beichman and Heng Li for help with PSMC analysis and Professor Eelco Rohling for assistance with interpretation of sea level data. We are grateful to Cisco Werner, Director of Scientific Programs and Chief Science Advisor for NOAA Fisheries, for funding sequencing of the vaquita genome. We thank Francoise Thibaud-Nissen of NCBI for annotation of the genome. Earlier versions of the manuscript have been improved thanks to careful review by Love Dalén, Mark Chaisson and four anonymous reviewers.

AUTHOR CONTRIBUTIONS

P.A.M., E.D.J., and O.A.R. initiated the project, and P.A.M., E.D.J., and O.F. designed and led research and analyses and cowrote the manuscript. F.I.A., B.H., J.R.B., J.M., and O.F. generpaated data. J.M., and S. Paez initiated the project for the VGP. A.P., A.R., B.H., A.F., G.F., K.H., J.R., J. Torrence, M.J.P.C., W.C., S. Palen and Y.V.B. contributed to data processing and genome assembly. M.L.H., A.C.M., J.A.F., and C.D.A. cultured cell lines, and A.W., B.L.T., C.R.S., F.M.D.G., J. Teilmann, L.R.-B., M.P.H.-J., R.S.W., S.S., T.R., and W.M. conducted the fieldwork to obtain and process the tissue samples. All authors contributed to interpretation of results and preparation of the manuscript.

DATA AVAILABILITY STATEMENT

The vaquita reference genome and all sequence data are available via the Vertebrate Genome Project GenomeArk website (https://vgp.github.io/genomeark/Phocoena_sinus/) and NCBI Genome database (Bioprojects PRJNA557831 and PRJNA557832). Annotation is available at NCBI (www.ncbi.nlm.nih.gov/genome/annotation_euk/Phocoena_sinus/100/). Ensembl annotation for the vaquita is available via the Ensembl data portal (ensembl.org/Phocoena_sinus/Info/Index).

ORCID

Phillip A. Morin https://orcid.org/0000-0002-3279-1519
Frederick I. Archer https://orcid.org/0000-0002-3179-4769
Jennifer R. Balacco https://orcid.org/0000-0001-7102-1632
Yury V. Bukhman https://orcid.org/0000-0002-8111-7651
William Chow https://orcid.org/0000-0002-9056-201X
Olivier Fedrigo https://orcid.org/0000-0002-6450-7551
Giulio Formenti https://orcid.org/0000-0002-7554-5991
Arkarachai Fungtammasan https://orcid.org/0000-0003-2398-0358
Frances M.D. Gulland https://orcid.org/0000-0002-6416-0156
Bettina Haase https://orcid.org/0000-0001-8945-7282
Mads Peter Heide-Jorgensen https://orcid.org/0000-0003-4846-7622
Kerstin Howe https://orcid.org/0000-0003-2237-513X
Jacquelyn Mountcastle https://orcid.org/0000-0003-1078-4905
Sarah Pelan https://orcid.org/0000-0001-8729-685X
Adam Phillippy https://orcid.org/0000-0003-2983-8934
Arang Rhie https://orcid.org/0000-0002-9809-8127
Jacqueline Robinson https://orcid.org/0000-0002-5556-815X

Oliver A. Ryder https://orcid.org/0000-0003-2427-763X
Barbara L. Taylor https://orcid.org/0000-0001-7620-0736
Jonas Teilmann https://orcid.org/0000-0002-4376-4700
James Torrance https://orcid.org/0000-0002-6117-8190
Randall S. Wells https://orcid.org/0000-0001-9793-4181
Andrew J. Westgate https://orcid.org/0000-0003-4340-7700
Erich D. Jarvis https://orcid.org/0000-0001-8931-5049

REFERENCES

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F. J., Lippman, Z. B., & Schatz, M. C. (2019). RaGOO: Fast and accurate reference-guided scaffolding of draft genomes. *Genome Biology*, *20*(1), 224. https://doi.org/10.1186/s13059-019-1829-6

Armstrong, E. E., Taylor, R. W., Prost, S., Blinston, P., van der Meer, E., Madzikanda, H., Mufute, O., Mandisodza-Chikerema, R., Stuelpnagel, J., Sillero-Zubiri, C., & Petrov, D. (2019). Cost-effective assembly of the African wild dog (*Lycaon pictus*) genome using linked reads. *Gigascience*, *8*(2), https://doi.org/10.1093/gigascience/giy124

Arnason, U., Lammers, F., Kumar, V., Nilsson, M. A., & Janke, A. (2018). Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Science. Advances*, *4*(4), eaap9873. https://doi.org/10.1126/sciadv.aap9873

Autenrieth, M., Hartmann, S., Lah, L., Roos, A., Dennis, A. B., & Tiedemann, R. (2018). High-quality whole-genome sequence of an abundant Holarctic odontocete, the harbour porpoise (*Phocoena phocoena*). *Molecular Ecology Resources*, *18*(6), 1469–1481. https://doi.org/10.1111/1755-0998.12932

Beichman, A. C., Phung, T. N., & Lohmueller, K. E. (2017). Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3*, *7*(11), 3605–3620. https://doi.org/10.1534/g3.117.300259

Ben Chehida, Y., Thumloup, J., Schumacher, C., Harkins, T., Aguilar, A., Borrel, A., Ferreira, M., Rojas-Bracho, L., Robertson, K. M., Taylor, B. L., & Fontaine, M. C. (2020). Mitochondrial genomics reveals evolutionary history of the porpoises (Phocoenidae) across the speciation continuum. *Scientific Reports*, *10*, 15190. https://doi.org/10.1101/851469

Brownell, R. L. Jr (1983). Phocoena sinus. *Mammalian Species*, *198*, 1–3. https://doi.org/10.2307/3503873

Brownell, R. L. Jr (1986). Distribution of the Vaquita, Phocoena-Sinus, in Mexican Waters. *Marine Mammal Science*, *2*, 299–305. https://doi.org/10.1111/j.1748-7692.1986.tb00137.x

Chattopadhyay, B., Garg, K. M., Yun Jing, S., Low, G. W., Frechette, J., & Rheindt, F. E. (2019). Conservation genomics in the fight to help the recovery of the critically endangered Siamese crocodile *Crocodylus siamensis*. *Molecular Ecology*, *28*(5), 936–950. https://doi.org/10.1111/mec.15023

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., Dunn, C., O'Malley, R., Figueroa-Balderas, R., Morales-Cruz, A., Cramer, G. R., Delledonne, M., Luo, C., Ecker, J. R., Cantu, D., Rank, D. R., & Schatz, M. C. (2016). Phased diploid genome assembly with single-molecule real-time sequencing. *Nature Methods*, *13*(12), 1050–1054. https://doi.org/10.1038/nmeth.4035

Chow, W., Brugger, K., Caccamo, M., Sealy, I., Torrance, J., & Howe, K. (2016). gEVAL – A web-based browser for evaluating genome assemblies. *Bioinformatics*, *32*(16), 2508–2510. https://doi.org/10.1093/bioinformatics/btw159

D'Agrosa, C., Lennert-Cody, C. E., & Vidal, O. (2000). Vaquita bycatch in Mexico's artisanal gillnet fisheries: Driving a small population to extinction. *Conservation Biology*, *14*(4), 1110–1119. https://doi.org/10.1046/j.1523-1739.2000.98191.x

Day, S. B., Bryant, E. H., & Meffert, L. M. (2003). The influence of variable rates of inbreeding on fitness, environmental responsiveness, and evolutionary potential. *Evolution*, *57*(6), 1314–1324. https://doi.org/10.1111/j.0014-3820.2003.tb00339.x

Dornburg, A., Brandley, M. C., McGowen, M. R., & Near, T. J. (2012). Relaxed clocks and inferences of heterogeneous patterns of nucleotide substitution and divergence time estimates across whales and dolphins (Mammalia: Cetacea). *Molecular Biology and Evolution*, *29*(2), 721–736. https://doi.org/10.1093/molbev/msr228

Excoffier, L., Dupanloup, I., Huerta-Sanchez, E., Sousa, V. C., & Foll, M. (2013). Robust demographic inference from genomic and SNP data. *PLOS Genetics*, *9*(10), e1003905. https://doi.org/10.1371/journal.pgen.1003905

Fan, G., Zhang, Y., Liu, X., Wang, J., Sun, Z., Sun, S., Zhang, H. E., Chen, J., Lv, M., Han, K., Tan, X., Hu, J., Guan, R., Fu, Y., Liu, S., Chen, X. I., Xu, Q., Qin, Y., Liu, L., … Liu, X. (2019). The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution. *Molecular Ecology Resources*, *19*(4), 944–956. https://doi.org/10.1111/1755-0998.13003

Foote, A. D., Liu, Y., Thomas, G. W. C., Vinař, T., Alföldi, J., Deng, J., Dugan, S., van Elk, C. E., Hunter, M. E., Joshi, V., Khan, Z., Kovar, C., Lee, S. L., Lindblad-Toh, K., Mancia, A., Nielsen, R., Qin, X., Qu, J., Raney, B. J., … Gibbs, R. A. (2015). Convergent evolution of the genomes of marine mammals. *Nature Genetics*, *47*(3), 272–275. https://doi.org/10.1038/ng.3198

Foote, A. D., Martin, M. D., Louis, M., Pacheco, G., Robertson, K. M., Sinding, M.-H., Amaral, A. R., Baird, R. W., Baker, C. S., Ballance, L., Barlow, J., Brownlow, A., Collins, T., Constantine, R., Dabin, W., Dalla Rosa, L., Davison, N. J., Durban, J. W., Esteban, R., … Morin, P. A. (2019). Killer whale genomes reveal a complex history of recurrent admixture and vicariance. *Molecular Ecology*, *28*, 3427–3444. https://doi.org/10.1111/mec.15099

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., Gibbs, R. A., Hanson, M. B., Korneliussen, T. S., Martin, M. D., Robertson, K. M., Sousa, V. C., Vieira, F. G., Vinar, T., Wade, P., Worley, K. C., Excoffier, L., Morin, P. A., Gilbert, M. T. P., & Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence in the killer whale. *Nature Communications*, *7*, 11693. https://doi.org/10.1038/ncomms11693

Formenti, G., Rhie, A., Balacco, J., Haase, B., Mountcastle, J., Fedrigo, O., Brown, S., Capodiferro, M., Al-Ajli, F. O., Ambrosini, R., Houde, P., Koren, S., Oliver, K., Smith, M., Skelton, J., Betteridge, E., Dolucan, J., Corton, C., Bista, I., … Jarvis, E. D. (2020). Complete and near error-free mitochondrial genome assemblies with long reads reveal unreported gene duplications and repeats. *BioRxiv*. https://www.biorxiv.org/content/10.1101/2020.06.30.177956v2

Garner, B. A., Hand, B. K., Amish, S. J., Bernatchez, L., Foster, J. T., Miller, K. M., Morin, P. A., Narum, S. R., O'Brien, S. J., Roffler, G., Templin, W. D., Sunnucks, P., Strait, J., Warheit, K. I., Seamons, T. R., Wenburg, J., Olsen, J., & Luikart, G. (2016). Genomics in conservation: Case studies and bridging the gap between data and application. *Trends in Ecology and Evolution*, *31*(2), 81–83. https://doi.org/10.1016/j.tree.2015.10.009

Garrison, E., & Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv:1207.3907 [q-bio.GN]*.

Ghurye, J., Pop, M., Koren, S., Bickhart, D., & Chin, C. S. (2017). Scaffolding of long read assemblies using long range contact information. *BMC Genomics*, *18*(1), 527. https://doi.org/10.1186/s12864-017-3879-z

Gilpin, M. E., & Soulé, M. E. (1986). Minimum viable populations: Processes of species extinction. In M. E. Soulé (Ed.), *Conservation biology: The science of scarcity and diversity* (pp. 19–34). Sinauer.

Harrisson, K. A., Pavlova, A., Telonis-Scott, M., & Sunnucks, P. (2014). Using genomics to characterize evolutionary potential for conservation of wild populations. *Evolutionary Applications*, *7*(9), 1008–1025. https://doi.org/10.1111/eva.12149

He, X., Johansson, M. L., & Heath, D. D. (2016). Role of genomics and transcriptomics in selection of reintroduction source populations. *Conservation Biology*, *30*(5), 1010–1018. https://doi.org/10.1111/cobi.12674

Hedrick, P. W., Robinson, J. A., Peterson, R. O., & Vucetich, J. A. (2019). Genetics and extinction and the example of Isle Royale wolves. *Animal Conservation*, *22*(3), 302–309. https://doi.org/10.1111/acv.12479

Hohn, A. A., Read, A. J., Fernandez, S., Vidal, O., & Findley, L. T. (1996). Life history of the vaquita, *Phocoena sinus* (Phocoenidae, Cetacea). *Journal of Zoology*, *239*, 235–251. https://doi.org/10.1111/j.1469-7998.1996.tb05450.x

Howe, K., Chow, W., Collins, J., Pelan, S., Pointon, D.-L., Sims, Y., … Wood, J. (2020). Significantly improving the quality of genome assemblies through curation. *bioRxiv*, https://doi.org/10.1101/2020.08.12.247734

Jaramillo-Legorreta, A. M., Cardenas-Hinojosa, G., Nieto-Garcia, E., Rojas-Bracho, L., Thomas, L., Ver Hoef, J. M., Moore, J., Taylor, B., Barlow, J., & Tregenza, N. (2019). Decline towards extinction of Mexico's vaquita porpoise (*Phocoena sinus*). *Royal Society Open Science*, *6*(7), 190598. https://doi.org/10.1098/rsos.190598

Jaramillo-Legorreta, A. M., Rojas-Bracho, L., & Gerrodette, T. (1999). A new abundance estimate for vaquitas: First step for recovery. *Marine Mammal Science*, *15*(4), 957–973. https://doi.org/10.1111/j.1748-7692.1999.tb00872.x

Keane, M., Semeiks, J., Webb, A. E., Li, Y. I., Quesada, V., Craig, T., Madsen, L. B., van Dam, S., Brawand, D., Marques, P. I., Michalak, P., Kang, L., Bhak, J., Yim, H.-S., Grishin, N. V., Nielsen, N. H., Heide-Jørgensen, M. P., Oziolor, E. M., Matson, C. W., … de Magalhães, J. P. (2015). Insights into the evolution of longevity from the bowhead whale genome. *Cell Reports*, *10*(1), 112–122. https://doi.org/10.1016/j.celrep.2014.12.008

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC Bioinformatics*, *15*, 356. https://doi.org/10.1186/s12859-014-0356-4

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*(2), R12. https://doi.org/10.1186/gb-2004-5-2-r12

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493–496. https://doi.org/10.1038/nature10231

Lischer, H. E. L., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, *18*(1), 474. https://doi.org/10.1186/s12859-017-1911-6

Mazet, O., Rodriguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: Instantaneous coalescence rates and human evolution–lessons for ancestral population size inference? *Heredity (Edinb)*, *116*(4), 362–371. https://doi.org/10.1038/hdy.2015.104

McGowen, M. R., Spaulding, M., & Gatesy, J. (2009). Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Molecular Phylogenetics and Evolution*, *53*(3), 891–906. https://doi.org/10.1016/j.ympev.2009.08.018

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Morin, P. A., Alexander, A., Blaxter, M., Caballero, S., Fedrigo, O., Fontaine, M. C., Foote, A. D., Kuraku, S., Maloney, B., McCarthy, M. L., McGowen, M. R., Mountcastle, J., Nery, M. F., Olsen, M. T., Rosel, P. E., & Jarvis, E. D. (2020). 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

Morin, P. A., Foote, A. D., Baker, C. S., Hancock-Hanser, B. L., Kaschner, K., Mate, B. R., Mesnick, S. L., Pease, V. L., Rosel, P. E., & Alexander, A. (2018). Demography or selection on linked cultural traits or genes? Investigating the driver of low mtDNA diversity in the sperm whale using complementary mitochondrial and nuclear genome analyses. *Molecular Ecology*, *27*(11), 2604–2619. https://doi.org/10.1111/mec.14698

Morin, P. A., Foote, A. D., Hill, C. M., Simon-Bouhet, B., Lang, A. R., & Louise, M. (2018). SNP discovery from single and multiplex genome assemblies of non-model organisms. In S. R. Head, P. Ordoukhanian, & D. Salomon (Eds.), *Next-generation sequencing. methods in molecular biology* (Vol. *1712*, pp. 113–144). Humana Press.

Munguia-Vega, A., Esquer-Garrigos, Y., Rojas-Bracho, L., Vazquez-Juarez, R., Castro-Prieto, A., & Flores-Ramirez, S. (2007). Genetic drift vs. natural selection in a long-term small isolated population: Major histocompatibility complex class II variation in the Gulf of California endemic porpoise (*Phocoena sinus*). *Molecular Ecology*, *16*(19), 4051–4065. https://doi.org/10.1111/j.1365-294X.2007.03319.x

Norris, K. S., & McFarland, W. N. (1958). A new harbor porpoise of the genus *Phocoena* from the Gulf of California. *Journal of Mammalogy*, *39*, 22–39. https://doi.org/10.2307/1376606

Orozco-terWengel, P. (2016). The devil is in the details: The effect of population structure on demographic inference. *Heredity*, *116*(4), 349–350. https://doi.org/10.1038/hdy.2016.9

Patil, A. B., Shinde, S. S., Raghavendra, S., Satish, B. N., Kushalappa, C. G., & Vijay, N. (2020). CoalQC – Quality control while inferring demographic histories from genomic data: Application to forest tree genomes. *bioRxiv*, https://doi.org/10.1101/2020.03.03.962365

Rhie, A., McCarthy, S. A., Fedrigo, O., Damas, J., Formenti, G., Koren, S., Uliano-Silva, M., Chow, W., Fungtammasan, A., Gedman, G.L., Cantin, L.J., Thibaud-Nissen, F., Haggerty, L., Lee, C., Ko, B.J., Kim, J., Bista, I., Smith, M., Haase, B., Mountcastle, J., … Jarvis, E. D. (2020). Towards complete and error-free genome assemblies of all vertebrate species. *bioRxiv*, https://doi.org/10.1101/2020.05.22.110833

Rhie, A., Walenz, B. P., Koren, S., & Phillippy, A. M. (2020). Merqury: Reference-free quality and phasing assessment for genome assemblies. *bioRxiv*, https://doi.org/10.1101/2020.03.15.992941

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1), 460. https://doi.org/10.1186/s12859-018-2485-7

Robinson, J. A., Brown, C., Kim, B. Y., Lohmueller, K. E., & Wayne, R. K. (2018). Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Current Biology*, *28*(21), 3487–3494 e3484. https://doi.org/10.1016/j.cub.2018.08.066

Robinson, J. A., Ortega-Del Vecchyo, D., Fan, Z., Kim, B. Y., vonHoldt, B. M., Marsden, C. D., Lohmueller, K. E., & Wayne, R. K. (2016). Genomic flatlining in the endangered island fox. *Current Biology*, *26*(9), 1183–1189. https://doi.org/10.1016/j.cub.2016.02.062

Robinson, J. A., Raikkonen, J., Vucetich, L. M., Vucetich, J. A., Peterson, R. O., Lohmueller, K. E., & Wayne, R. K. (2019). Genomic signatures of extensive inbreeding in Isle Royale wolves, a population on the threshold of extinction. *Science Advances*, *5*(5), eaau0757. https://doi.org/10.1126/sciadv.aau0757

Rodriguez-Perez, M. Y., Aurioles-Gamboa, D., Sanchez-Velasco, L., Lavin, M. F., & Newsome, S. D. (2018). Identifying critical habitat of the endangered vaquita (Phocoena sinus) with regional delta C-13 and delta N-15 isoscapes of the Upper Gulf of California, Mexico. *Marine Mammal Science*, *34*(3), 790–805. https://doi.org/10.1111/mms.12483

Rohling, E. J., Hibbert, F. D., Williams, F. H., Grant, K. M., Marino, G., Foster, G. L., Hennekam, R., de Lange, G. J., Roberts, A. P., Yu, J. M., Webster, J. M., & Yokoyama, Y. (2017). Differences between the last

two glacial maxima and implications for ice-sheet, delta O-18, and sea-level reconstructions. *Quaternary Science Reviews*, *176*, 1–28. https://doi.org/10.1016/j.quascirev.2017.09.009

Rojas-Bracho, L., Gulland, F., Smith, C. R., Taylor, B., Wells, R. S., Thomas, P. O., Bauer, B., Heide-Jørgensen, M. P., Teilmann, J., Dietz, R., Balle, J. D., Jensen, M. V., Sinding, M., Jaramillo-Legorreta, A., Abel, G., Read, A. J., Westgate, A. J., Colegrove, K., Gomez, F., … Walker, S. (2019). A field effort to capture critically endangered vaquitas *Phocoena sinus* for protection from entanglement in illegal gillnets. *Endangered Species Research*, *38*, 11–27. https://doi.org/10.3354/esr00931

Rojas-Bracho, L., & Reeves, R. R. (2013). Vaquitas and gillnets: Mexico's ultimate cetacean conservation challenge. *Endangered Species Research*, *21*(1), 77–87. https://doi.org/10.3354/esr00501

Rojas-Bracho, L., & Taylor, B. L. (1999). Risk factors affecting the vaquita (*Phocoena sinus*). *Marine Mammal Science*, *15*(4), 974–989. https://doi.org/10.1111/j.1748-7692.1999.tb00873.x

Rosel, P. E., Haygood, M. G., & Perrin, W. F. (1995). Phylogenetic relationships among the true porpoises (Cetacea:Phocoenidae). *Molecular Phylogenetics and Evolution*, *4*(4), 463–474. https://doi.org/10.1006/mpev.1995.1043

Rosel, P. E., & Rojas-Bracho, L. (1999). Mitochondrial DNA variation in the critically endangered vaquita *Phocoena sinus* Norris and Macfarland, 1958. *Marine Mammal Science*, *15*(4), 990–1003. https://doi.org/10.1111/j.1748-7692.1999.tb00874.x

Siddall, M., Rohling, E. J., Almogi-Labin, A., Hemleben, C., Meischner, D., Schmelzer, I., & Smeed, D. A. (2003). Sea-level fluctuations during the last glacial cycle. *Nature*, *423*(6942), 853–858. https://doi.org/10.1038/nature01690

Smit, A., Hubley, R., & Green, P. (2013–2015). *RepeatMasker Open 4.0*. Retrieved from http://www.repeatmasker.org

Springer, M. S., Emerling, C. A., Fugate, N., Patel, R., Starrett, J., Morin, P. A., Hayashi, C., & Gatesy, J. (2016). Inactivation of cone-specific phototransduction genes in rod monochromatic cetaceans. *Frontiers in Ecology and Evolution*, *4*, 61. https://doi.org/10.3389/fevo.2016.00061

Springer, M. S., Starrett, J., Morin, P. A., Lanzetti, A., Hayashi, C., & Gatesy, J. (2016). Inactivation of C4orf26 in toothless placental mammals. *Molecular Phylogenetics and Evolution*, *95*, 34–45. https://doi.org/10.1016/j.ympev.2015.11.002

Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. *Genome Biology*, *19*(131), https://doi.org/10.1186/s13059-018-1520-3

Taylor, B. L., Chivers, S. J., Larese, J., & Perrin, W. F. (2007). *Generation length and percent mature estimates for IUCN assessments of cetaceans* (Administrative Report LJ-07-01). Retrieved from https://swfsc.noaa.gov/publications/TM/SWFSC/NOAA-TM-NMFS-SWFSC-550.pdf

Taylor, B. L., & Rojas-Bracho, L. (1999). Examining the risk of inbreeding depression in a naturally rare cetacean, the vaquita (*Phocoena sinus*). *Marine Mammal Science*, *15*(4), 1004–1028. https://doi.org/10.1111/j.1748-7692.1999.tb00875.x

Taylor, B. L., Wells, R. S., Olson, P. A., Brownell, R. L., Gulland, F. M. D., Read, A. J., … Rojas-Bracho, L. (2019). Likely annual calving in the vaquita, *Phocoena sinus*: A new hope? *Marine Mammal Science*, *35*(4), 1603–1612. https://doi.org/10.1111/mms.12595

Thomas, L., Jaramillo-Legorreta, A., Cardenas-Hinojosa, G., Nieto-Garcia, E., Rojas-Bracho, L., Ver Hoef, J. M., Moore, J., Taylor, B., Barlow, J., & Tregenza, N. (2017). Last call: Passive acoustic monitoring shows continued rapid decline of critically endangered vaquita. *Journal of the Acoustical Society of America*, *142*(5), El512–El517. https://doi.org/10.1121/1.5011673

Tollis, M., Robbins, J., Webb, A. E., Kuderna, L. F. K., Caulin, A. F., Garcia, J. D., Bèrubè, M., Pourmand, N., Marques-Bonet, T., O'Connell, M. J., Palsbøll, P. J., & Maley, C. C. (2019). 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

Tunstall, T., Kock, R., Vahala, J., Diekhans, M., Fiddes, I., Armstrong, J., Paten, B., Ryder, O. A., & Steiner, C. C. (2018). Evaluating recovery potential of the northern white rhinoceros from cryopreserved somatic cells. *Genome Research*, *28*(6), 780–788. https://doi.org/10.1101/gr.227603.117

Turvey, S. T., Pitman, R. L., Taylor, B. L., Barlow, J., Akamatsu, T., Barrett, L. A., Zhao, X., Reeves, R. R., Stewart, B. S., Wang, K., Wei, Z., Zhang, X., Pusser, L. T., Richlen, M., Brandon, J. R., & Wang, D. (2007). First human-caused extinction of a cetacean species? *Biology Letters*, *3*, 537–540. https://doi.org/10.1098/rsbl.2007.0292

Van Cise, A. M., Baird, R. W., Baker, C. S., Cerchio, S., Claridge, D., Fielding, R., Hancock-Hanser, B., Marrero, J., Martien, K. K., Mignucci-Giannoni, A. A., Oleson, E. M., Oremus, M., Poole, M. M., Rosel, P. E., Taylor, B. L., & Morin, P. A. (2019). Oceanographic barriers, divergence, and admixture: Phylogeography and taxonomy of two putative subspecies of short-finned pilot whale. *Molecular Ecology*, *28*, 2886–2902. https://doi.org/10.1111/mec.15107

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2017). BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319

Westbury, M. V., Hartmann, S., Barlow, A., Wiesel, I., Leo, V., Welch, R., Parker, D. M., Sicks, F., Ludwig, A., Dalén, L., & Hofreiter, M. (2018). Extended and continuous decline in effective population size results in low genomic diversity in the world's rarest hyena species, the brown hyena. *Molecular Biology & Evolution*, *35*(5), 1225–1237. https://doi.org/10.1093/molbev/msy037

Westbury, M. V., Petersen, B., Garde, E., Heide-Jorgensen, M. P., & Lorenzen, E. D. (2019). Narwhal genome reveals long-term low genetic diversity despite current large abundance size. *iScience*, *15*, 592–599. https://doi.org/10.1016/j.isci.2019.03.023

Xue, Y., Prado-Martinez, J., Sudmant, P. H., Narasimhan, V., Ayub, Q., Szpak, M., Frandsen, P., Chen, Y., Yngvadottir, B., Cooper, D. N., de Manuel, M., Hernandez-Rodriguez, J., Lobon, I., Siegismund, H. R., Pagani, L., Quail, M. A., Hvilsom, C., Mudakikwa, A., Eichler, E. E., … Scally, A. (2015). Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science*, *348*(6231), 242–245. https://doi.org/10.1126/science.aaa3952

Yim, H.-S., Cho, Y. S., Guang, X., Kang, S. G., Jeong, J.-Y., Cha, S.-S., Oh, H.-M., Lee, J.-H., Yang, E. C., Kwon, K. K., Kim, Y. J., Kim, T. W., Kim, W., Jeon, J. H., Kim, S.-J., Choi, D. H., Jho, S., Kim, H.-M., Ko, J., … Lee, J.-H. (2014). Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics*, *46*, 88–92. https://doi.org/10.1038/ng.2835

Zhou, X., Guang, X., Sun, D. I., Xu, S., Li, M., Seim, I., Jie, W., Yang, L., Zhu, Q., Xu, J., Gao, Q., Kaya, A., Dou, Q., Chen, B., Ren, W., Li, S., Zhou, K., Gladyshev, V. N., Nielsen, R., … Yang, G. (2018). Population genomics of finless porpoises reveal an incipient cetacean species adapted to freshwater. *Nature Communications*, *9*(1), 1276. https://doi.org/10.1038/s41467-018-03722-x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Morin PA, Archer FI, Avila CD, et al. Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Mol Ecol Resour*. 2021;21:1008–1020. https://doi.org/10.1111/1755-0998.13284