Molecular Ecology Resources

WILEY

MOLECULAR ECOLOGY RESOURCES

RESOURCE ARTICLE OPEN ACCESS

Breath to Genome: Whole Genome Sequencing of Large Whales From Blow Sampling

Éadin N. O'Mahony[1,2,3,4] | Morgan L. McCarthy[3,5] | Eric M. Keen[4,6] | Janie Wray[4] | Christie J. McMillan[7] | Sheila J. Thornton[8] | Luke E. Rendell[1,2] | Oscar E. Gaggiotti[2] | Morten Tange Olsen[3]

[1]Sea Mammal Research Unit, Scottish Oceans Institute, University of St Andrews, St Andrews, UK | [2]Centre for Biological Diversity, University of St Andrews, St Andrews, UK | [3]Section for Molecular Ecology & Evolution, Globe Institute, University of Copenhagen, Copenhagen, Denmark | [4]North Coast Cetacean Society, Alert Bay, British Columbia, Canada | [5]Department of Population Analysis and Monitoring, Swedish Museum of Natural History, Stockholm, Sweden | [6]Sewanee: The University of the South, Sewanee, Tennessee, USA | [7]Pacific Biological Station, Fisheries and Oceans Canada, Nanaimo, British Columbia, Canada | [8]Pacific Science Enterprise Centre, Fisheries and Oceans Canada, West Vancouver, British Columbia, Canada

Correspondence: Éadin N. O'Mahony (enao1@st-andrews.ac.uk)

Received: 9 February 2026 | **Revised:** 30 September 2026 | **Accepted:** 8 October 2026

Keywords: baleen whales | blow sampling | low coverage whole genomes | *Megaptera novaeangliae* | shotgun sequencing | SNPs

ABSTRACT

Preserving biodiversity requires the genetic monitoring of wild populations, which is traditionally done through invasive sampling techniques for DNA collection. However, this approach can impact animal welfare or be impossible to implement due to behavioural responses to anthropogenic factors or national/international legislation. For cetaceans, a promising non-invasive alternative approach is the collection of exhaled respiratory vapour, or 'blow', to generate individual genetic profiles. Here, we demonstrate the feasibility of generating whole nuclear and mitochondrial genomes from the blow of humpback whales (*Megaptera novaeangliae*) with data of sufficient quality for population genomics. A total of 58 blow samples from 26 Northeast Pacific humpback whales were collected in Gitga'at First Nation territory using a commercially available drone. The high endogenous DNA content at 84% on average allowed us to generate low-coverage nuclear genomes (mean 2.3×, range 0.2× to 3.5×) and high coverage mitochondrial genomes (mean 94×, range 7.7× to 151.3×) for inference of population structure, diversity and phylogenomics. The reliability of the blow-derived genomes was demonstrated by direct comparison between replicate blow samples, as well as paired tissue samples from a subset of individuals. Although applicability will likely vary across species depending on factors such as elusiveness, body size and surface time, this approach represents a major step towards non-invasive cetacean sampling that can yield not only host genomes but also microbiome and hormone profiles. As such, our approach opens new avenues for conservation genomics, population monitoring and microbiome-based health assessments in the wild and can fill critical data gaps in cetacean genomic databases.

1 | Introduction

Genomic data has become an important tool in conservation biology with the ability to accurately profile genetic population structure, measure relatedness and diversity, estimate effective population sizes and much more (Cammen et al. 2016). The best genomic data are typically obtained from invasive tissue sampling, but for species of perilous conservation status, such sampling can be undesirable and restricted for management and ethical concerns (Zemanova 2020). Non-invasive methods of collecting genetic samples can be a preferred alternative to avoid harm and minimise disturbance to individuals,

Oscar E. Gaggiotti and Morten Tange Olsen shared anchor-authorship.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2026 The Author(s). *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

an increasingly pertinent consideration in an era of persistent, pervasive anthropogenic pressures on nature (Byrne et al. 2021; Zemanova 2019). In addition to ethical and legal considerations, non-invasive methods often allow sampling when and/or where more conventional sampling methods (e.g., biopsies) are not possible, such as in species that are elusive or protected from harm or take (e.g., snow paw print sampling of polar bears *Ursus maritimus*; Von Duyke et al. 2023). Finally, physical access to the animal and its habitat can also be challenging, as is the case for many marine species (Visser et al. 2021). Fortunately, minimally- and non-invasive genetic sampling methods and their ethical benefits are becoming increasingly available thanks to new sampling technologies and the rapid advances and decreasing costs of genome sequencing technologies (Carroll et al. 2018; Waits and Paetkau 2005).

Minimally- or non-invasive genetic sampling of marine species varies between taxonomic groups of interest, with environmental (e)DNA often providing a broad-scale overview of entire communities (e.g., Port et al. 2016), while more species-specific approaches range from mucus swabs in elasmobranchs (e.g., Kashiwagi et al. 2015; Lieber et al. 2013) to cetacean fluke-footprint eDNA (e.g., Székely et al. 2021), sloughed skin (e.g., Amos et al. 1992; Neveceralova et al. 2022) and exhaled respiratory vapour or 'blow' sampling (e.g., Frère et al. 2010). In the context of cetacean blow sampling, drones are emerging as the method of choice, having been applied at various latitudes to a range of mysticete and odontocete species, including humpback whales *Megaptera novaeangliae*, fin whales *Balaenoptera physalus*, blue whales *B. musculus*, grey whales *Eschrichtius robustus*, sperm whales *Physeter macrocephalus* and bottlenose dolphins *Tursiops truncatus* (Acevedo-Whitehouse et al. 2010; Atkinson et al. 2021; Centelleghe et al. 2021; Costa et al. 2022; O'Mahony et al. 2024). The blow, whether collected with a drone or an extended pole, has been used to assess whale health through the characterisation of respiratory microbiota (e.g., Acevedo-Whitehouse et al. 2010) and hormonal profiles (e.g., Burgess et al. 2018; Hogg et al. 2009). With varying levels of success, host DNA has been extracted from the blow of a range of species, all of which amplify short fragments of the mitochondrial genome and/or a small number of microsatellite loci from the nuclear genome (Atkinson et al. 2021; Borowska et al. 2014; Frère et al. 2010; Neveceralova et al. 2025; O'Mahony et al. 2024; Richard et al. 2017; Robinson and Nuuttila 2020). However, these molecular markers can have limited power at detecting population structure and inference of diversity, relatedness, inbreeding, genomic erosion, adaptive genetic variation and demographic histories, amongst others, information that can inform the delineation of management units and the setting of meaningful conservation targets (Funk et al. 2012; Hohenlohe et al. 2021; Montana et al. 2024; Tokarska et al. 2009).

The development of high-throughput sequencing (HTS) alongside a reduction in costs and increase in analytical tools has seen a shift from working with a single or small numbers of molecular markers, such as mitochondrial control region haplotypes and microsatellites, to whole nuclear and mitochondrial genome analyses (Allendorf et al. 2010). Additionally, these rapidly advancing molecular techniques allow for progressively lower quantities of sample DNA to generate whole genomes (de Flamingh et al. 2022), and analytical approaches such as the inference of genotype likelihoods from genomes at low depth of coverage, allow for highly reliable population genomic inference (Fuentes-Pardo and Ruzzante 2017; Korneliussen et al. 2014; Lou et al. 2021).

Here, we demonstrate for the first time the feasibility of obtaining low-coverage nuclear genomes and high coverage mitochondrial genomes (hereafter 'mitogenomes') from the blow of free ranging humpback whales to generate data of sufficient quality for population genomics. Samples were collected from humpback whales in Gitga'at First Nation territory in what is now known as British Columbia (BC), Canada, as described by our previous work detailed in O'Mahony et al. (2024), using a small commercially available unoccupied aerial vehicle (UAV), or drone. This method significantly reduces the disturbance to the target animals relative to current methods by removing the need for close vessel-approaches and invasive biopsy sampling (Christiansen et al. 2020).

2 | Materials and Methods

2.1 | Method Validation and Population Genomic Inference

Demonstrating the reliability of humpback whale blow samples to generate whole genome sequencing (WGS) data requires several analytical steps. First, we visually confirmed the identity of sampled animals using photographic identification (ID). Second, we assessed whether blow sample sequencing depth and endogenous content allow for reliable inference of population genomic metrics such as relatedness, population structure, heterozygosity and runs of homozygosity (ROH) relative to those calculated from traditional biopsy samples. Third, we demonstrate that our blow samples are not prone to contamination with DNA from other animals in the vicinity, as this could confuse individual recognition, sex-determination, relatedness inference and inflate genetic diversity metrics. To this end, duplicate blow samples collected separately from the same field-identified individuals were analysed to confirm that they are true genomic replicates of each other. The low-coverage WGS from the blow of two individual humpback whales were also compared to medium coverage WGS from biopsy samples visually matched to the same individuals (15× to 16× depth of coverage). Finally, to demonstrate the usefulness of our blow genomes in a population genomic context, we combined our data with published data to infer the diversity and structure of humpback whales in the North Pacific and North Atlantic.

2.2 | Sample Collection

2.2.1 | Blow Sampling

We used a drone to collect 58 blow samples from 26 unique field-identified humpback whales between July and September of 2022 and 2023 in the Kitimat Fjord System (Gitga'at First Nation territory) of northern BC, Canada, through research agreements between the Gitga'at First Nation and the North Coast Cetacean Society (Box 1; Figure 1A,B). Samples were collected on a sterile petri dish attached by custom 3D-printed

housing to a small quadcopter drone (DJI Mavic 2 Pro) operated by author ÉOM, a Transport Canada Advanced Operations drone pilot (under Department of Fisheries and Oceans research permit #XMMS12022; Figure 1C). Immediately upon sample collection, the drone was returned to the operator and petri dishes sealed (Figure 1D) and placed on ice until return to the field station. Petri dishes were then rinsed with 0.5 mL Longmire's solution and 1.5 mL distilled H_2O using a sterile scraper to ensure all of the blow droplets were washed into solution. This solution was then pipetted into a 2 mL cryovial and frozen in a household freezer set to −20°C until being shipped on dry ice to the laboratories of the Globe Institute, University of Copenhagen, Denmark.

BOX 1 | Positionality and reflexivity statement.

We are an international group of marine mammal researchers united by the motivation to simultaneously minimise impact on the cetacean populations we study, while maximising the research output possible from the data we collect. We have an almost equal gender balance, and are collaborating across community-, NGO-, governmental- and academic-rooted organisations. Authors É.N.O., M.L.M., L.E.R., O.E.G. and M.T.O. are based at European universities and museums, whilst E.M.K. is based at a university in the United States. Authors J.W., C.J.M. and S.J.T. are based in Canada. This study formed part of author É.N.O's Doctoral thesis (O'Mahony 2026). All authors are grateful for our ongoing collaboration with multiple First Nation communities, particularly the Gitga'at First Nation and extend our gratitude for the opportunity to study the recovering populations of baleen whales foraging in the fjords of Gitga'at territory in the eastern North Pacific.

Blow samples were collected opportunistically in good weather conditions (Beaufort < 3). The most upwind whale of a group was targeted to minimise the risk of intraspecific contamination and the drone pilot used distinct body features, such as visible scars, to verify individual identity via the live video feed, ensuring only one individual was sampled per flight. For each flight, the number of exhalations sampled was maximised from the same surfacing event of the same individual, with the drone pilot and research assistant maintaining visual contact of the target whale (1–5 breaths sampled per surfacing event). Individual whales were typically only sampled once per encounter; however, duplicate and triplicate samples were collected from individuals during different encounters to allow assessment of sample contamination and reliability of genomic data across different group sizes and behavioural states. To ensure minimal invasiveness, resting/sleeping whales were not sampled and the behavioural responses of the sampled whale were monitored in real-time by the drone pilot. A response scale adapted from Weinrich et al. (1992), which we have used for blow sampling since 2019 (O'Mahony et al. 2024), was also applied here to grade levels

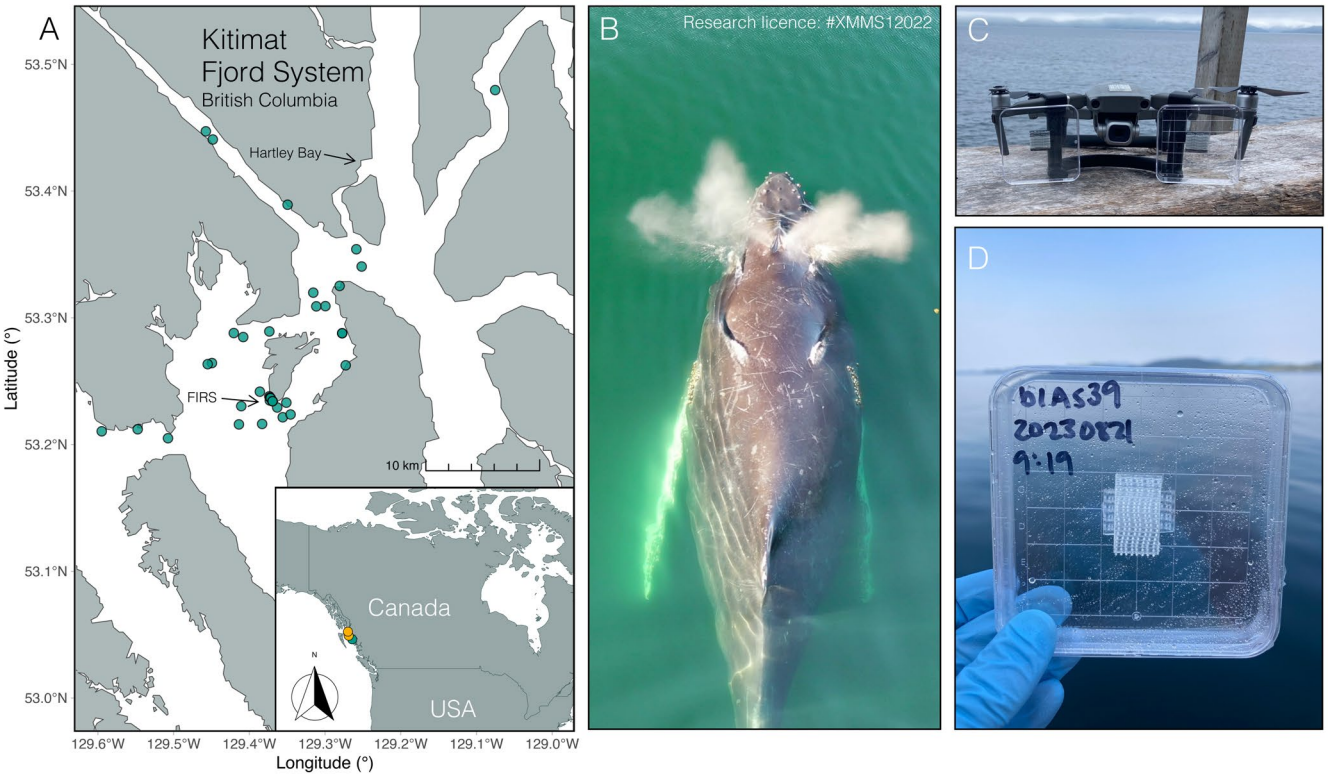


FIGURE 1 | (A) The blow sampling study area, which falls within Gitga'at First Nation territory, British Columbia, Canada. Blow sampling locations are plotted in green and biopsy sampling locations in yellow. The research station is marked as 'FIRS' (Fin Island Research Station) on the southern end of Fin Island. (B) A frame extracted from a blow sampling flight to demonstrate the feasibility of using scarring visible on the back of the whale to cross-reference the individual ID between replicate sampling events. (C) The drone with custom 3D-printed attachments to hold the sterile petri dish secure shortly before each flight. (D) An example of humpback whale blow samples shortly after collection, before processing.

of behavioural response of the target whales to the drone and/or our presence in the research vessel, with flights being terminated if a response was detected. Blow sampling was reviewed and approved by the University of St Andrews' Animal Welfare and Ethics Review Body (Reference #BL16194).

2.2.2 | Biopsy Sampling

To validate the blow sampling approach, two biopsy samples were collected in 2022 by Fisheries and Oceans Canada (DFO) under Marine Mammal Licence Number MML-01. Individual whales were identified prior to sampling using standard fluke-ID photographs and techniques. Samples were collected using a sterile stainless-steel biopsy tip with a cutting edge (40 mm length, 8 mm diameter) affixed to a carbon fibre bolt with a float positioned anterior to the biopsy tip. Bolts were deployed from a Barnett recurve crossbow with a draw weight of 150 lbs. The biopsy tips were cleaned and sterilised between each use following protocols approved by the Pacific Region Animal Care Committee. Samples were placed in a cooler on ice immediately after collection and transferred to liquid nitrogen at the end of the day. Upon return from the field, samples were stored in a −80°C freezer until further processing.

2.2.3 | Photo-Identification of Blow-Sampled Animals

Correct field IDs can be challenging to obtain in certain circumstances, such as when there is a large group of whales foraging together or if an individual whale does not fluke. The use of a drone for blow sampling, which allows for the simultaneous recording of the flight in 4K video, presents the opportunity to recheck field IDs of sampled whales using the distinct markings on the rostra and dorsal regions of the whales, even years after the sampling event (Figure 1B; and see Degollada et al. 2023; O'Mahony et al. 2024). Individual identification was performed by the standard paired fluke photographic approach (described in O'Mahony et al. 2024) using 4K video from the drone as well as by DSLR cameras from the observer platform. Fluke identification photographs were manually matched against the North Coast Cetacean Society's long-term humpback whale fluke catalogue (spanning the years 2001 to 2023). Drone video footage of each sampling flight was manually watched to: (1) extract any fluke identifications, then used to manually corroborate photographic IDs; (2) confirm through rostrum markings that the sampled whale was the same throughout the sampling event; and (3) count the number of exhales flown through with the drone to assess the relationship between the number of breaths collected and sequencing depth of coverage. For this, we used a linear regression model with number of breaths as the explanatory variable, depth of coverage as the response variable and examination of residual diagnostics to confirm that model assumptions were met.

2.3 | DNA Extraction and Sequencing

DNA was extracted from the 58 blow samples using a slightly modified Qiagen DNeasy Blood and Tissue extraction protocol (see Supporting Information and O'Mahony et al. 2024). The DNA concentration of the samples was quantified with the Qubit High-Sensitivity assay. Samples were extracted with negative controls (extraction blanks) and two positive controls—both ringed seals *Pusa hispida*, sampled in Pond Inlet and Resolute, Nunavut, Canada. Humpback whale tissue samples were extracted using a phenol–chloroform extraction with proteinase K digestion (see Supporting Information for further detail). Library preparation and shotgun sequencing of blow samples was performed by BGI China using 20 μL of sample extracts. Blow sample extracts were prepared for sequencing using the KAPA HyperPrep library preparation kit for low input DNA and shotgun sequenced (paired-end 100 bp) on the DNBSEQ platform producing 10 Gb of data per sample. Biopsies were shotgun sequenced by Genome Quebec (Montreal, Canada) using a PCR-free protocol on Illumina NovaSeq 6000 S4 with paired end 150 bp reads to a target depth of 15× to 20×.

2.4 | Read Processing and Alignment

The WGS raw read data of 58 humpback whale blow samples and two biopsy-collected samples were quality controlled using FastQC (Andrews 2010) and then cleaned using fastp v.0.23.4 (Chen et al. 2018) with the default settings, before being mapped to the humpback whale chromosome-level nuclear genome (sequenced from a high quality tissue sample taken from a calf on the Hawaiian breeding ground of the study population's migratory corridor; Carminati et al. 2024) and separately mapped to the mitochondrial genome of the same individual (GenBank assembly GCA_041834305.1; Carminati et al. 2024) using BWA v.0.7.17 with the MEM algorithm (Li 2013). Alignments were sorted, converted into BAM format and exact duplicate single-end reads (usually PCR duplicates) were removed using SAMtools v.1.17 (rmdup -S) (Li et al. 2009). The endogenous content of humpback whale DNA within the blow samples was calculated as the fraction of reads mapping to the humpback whale reference genome to the total number of reads per sample. The depth of coverage (as the mean number of reads mapping to each position along the complete genome) was calculated using SAMtools (Li et al. 2009). A bed file of genomic interspersed repeats (i.e., regions of highly repetitive DNA sequences) within the humpback whale genome was created using RepeatMasker v.4.1.5 (Smit et al. 2013) and then reads mapping to these regions were masked from the cleaned bam files using BEDTools v.2.31.0 (Quinlan and Hall 2010). We serially down sampled biopsy genomes to evaluate the effect of depth of coverage on heterozygosity and ROH using seqtk v.1.4 (Li 2017; for further detail see Supporting Information).

Publicly available humpback whale WGS data ($n = 9$) were mapped to the same reference genome as the data analysed in this study. These included a biopsy from the western North Atlantic (8× depth; accession SRR8386009; Tollis et al. 2019), a sample from cell culture (19× depth; accession SRR5665639; Árnason et al. 2018) and six unrelated North Atlantic individuals from biopsy (26× depth on average; accessions SRR25114440-25114446; Suárez-Menéndez et al. 2023), a North Pacific sample sequenced by DNA Zoo (62× depth; accession SRR17854487), and the two biopsy samples presented in this study. Raw read data were processed with the same steps as used for the blow sample dataset.

2.5 | Library Complexity

Blow sample BAM files that had been mapped (*bwa mem*) and sorted (*samtools view* and *samtools sort*), but not yet deduplicated, were used to assess library complexity with preseq v.3.1.2 (Daley and Smith 2013). Preseq predicts the yield of distinct read pairs for varying quantities of total reads sequenced, both for smaller experiments (*preseq c_curve*) and the expected yield for larger future experiments (*preseq lc_extrap*). BAM files were pre-filtered to properly paired, primary alignments (*samtools view* -f 2 -F 2304) and then preseq *c_curve* and *lc_extrap* were run specifying the -P flag for paired-end reads. For visualisation with ggplot2, library complexity extrapolations were truncated to five times each sample's observed number of reads, as a proxy for five-times increased sequencing effort (Figure S2). We report the predicted fold increase in distinct read pairs from doubling sequencing effort with samples below 1× depth of coverage removed, as these samples theoretically would have been discarded from further sequencing due to poor quality. Further, the preseq function *gc_extrap* was used to predict genomic breadth of coverage with greater sequencing effort and both current breadth of coverage (at 10 Gb per sample) and extrapolated breadth of coverage for doubling sequencing effort (20 Gb per sample) are reported.

2.6 | Mitochondrial Genomes

A *Megaptera novaeangliae* mitogenome from the same reference genome used for the WGS data analyses was downloaded from NCBI (GenBank accession PP475430). Prior to mapping, a 70 bp pad was added to each end of the linearized mitogenome to ensure sufficient mappability across the artificial break of the circular mitogenome. The reference mitogenome was first indexed using SAMtools v.1.17 *faidx* (Li et al. 2009) and BWA v.0.7.17 *index* (Li 2013) and blow sample, biopsy and publicly available raw read data were mapped using BWA *mem* (Li 2013). Duplicates were removed (*samtools rmdup*) and once the depth (+/− one standard deviation) was calculated for each sample, a consensus mitogenome was called using ANGSD v.0.940-2 (-doFasta 2, calling the most common base at each site across all reads; Korneliussen et al. 2014).

Consensus mitogenomes were imported into Geneious Prime v.2025.0.3 and a multiple alignment was performed using the MAFFT v.7.490 alignment algorithm with default settings (Katoh et al. 2002; Katoh and Standley 2013). We included publicly available humpback whale whole mitogenomes ($n = 10$; GenBank accession numbers Table S4) and two North Atlantic fin whale mitogenomes as an outgroup (GenBank accession numbers: NC_001321 (Valverde et al. 1994) and OZ239533, submitted by the Darwin Tree of Life Project). Mitochondrial control region haplotypes were assigned according to the haplotypes described in Baker et al. (2013) and haplotype frequencies calculated for comparison to frequencies found in the northern BC region of this study. The post-QC subset of samples, plus sample MN-b-070 (because all control region haplotypes were then represented by at least one sequence) and the publicly available mitogenomes were exported as a Phylip alignment and a maximum likelihood phylogenetic tree constructed using IQ-TREE (Nguyen et al. 2015) with branch support calculated using 100 replicate bootstraps. Figures 3B and S6 were created using FigTree v.1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

2.7 | Nuclear Genome Data Quality Control

Several quality control (QC) steps were taken to filter first the blow sample dataset with the biopsy-sourced data and later the 'merged dataset', where 25 merged samples were created from replicate blow samples. Specifically, for the 58 unmerged samples: (1) samples with less than 1× depth of coverage were removed. (2) Of the known replicate sets, the highest depth of coverage sample was retained. (3) Pairwise relatedness, as the KING-robust kinship coefficient (Manichaikul et al. 2010), was calculated for all remaining samples (see details in section *Relatedness*) and of the related pairs with KING > 0.125, the sample of higher depth of coverage was retained. (4) Finally, we used the 'perfect-sample approach' (Orlando et al. 2013) to determine if any remaining samples had relatively elevated error rates (expanded in section *Relative error rates*). We then calculated the upper 95th percentile and removed any samples falling above this threshold.

For the merged dataset of replicate samples from the same individual humpback whales the QC steps were: (1) we assigned mitochondrial control region haplotypes according to haplotypes used by Baker et al. (2013). Based on the field-assigned IDs, we examined pairs of replicate samples, and for any pairs with mismatching haplotypes we re-assessed ID photographs and watched the associated drone footage to confirm if the pair was in fact a mismatch based on scarring on the back of the animal (Figure 1B). (2) In parallel to this, we sorted estimates of pairwise relatedness (see section *Relatedness* below) according to whether they are replicates (1) or not (0). Outlier replicate pairs falling below the 25th percentile threshold had their IDs re-assessed. Any pairs with confirmed mis-assigned IDs were flagged and not used to create the merged dataset. (3) Remaining replicate sets were merged using SAMtools v.1.17 *merge* to increase the depth of coverage for individual humpback whales. (4) Pairwise relatedness was recalculated using this merged dataset to flag any highly related pairs. Of pairs with a KING > 0.125, the higher depth of coverage individual was retained. (5) Finally, the 'perfect sample approach' was also applied to the merged samples using the same perfect individual and outgroup. The upper 95th percentile was once again used as the threshold over which samples should be excluded.

2.8 | Relative Error Rates

To identify samples with relatively high error rates, perhaps stemming from low depth of coverage or contamination, we used ANGSD (Korneliussen et al. 2014) to measure the distance from a 'perfect sample' to an outgroup and used this to quantify deviations in the frequency of derived alleles in each sample relative to the perfect sample. We used the consensus fasta of the fin whale DNA Zoo reference genome (SRR16970344; Dudchenko et al. 2017, 2018) mapped to the humpback reference as the outgroup (-anc) and the highest depth of coverage North Pacific whole genome (SRR17854487; Dudchenko et al. 2017,

2018) as the perfect individual (-ref). We generated consensus fasta sequences using the ANGSD option -doFasta 2, which samples the most common base. For the error rate estimation, we followed standard thresholds—using a randomly selected base with -doAncError 1, a mapping and base qualities threshold of 30 and using only autosomes (minMapQ 30 -minQ 30 -rf ${autosomes}).

2.9 | Relatedness

We estimated pairwise relatedness amongst all possible pairs of blow samples and biopsies (including biopsies subsampled to lower depths of coverage) using no minimum depth threshold, with the goal of comparing relatedness coefficients between pairs of field replicates (i.e., same assigned field ID, even if sampled on different encounters) with those between pairs of non-replicate samples. This demonstrates the reliability of blow-sourced genomes, given that the sequenced genomes sampled from the same individual during different field encounters remains consistent. Estimation of relatedness also formed a QC step, as mentioned above we were able to identify incorrect IDs made in the field by lower than expected pairwise relatedness coefficients between blow samples that were identified as replicates.

We used NgsRelate v2.0 (Korneliussen and Moltke 2015) to calculate KING-robust kinship coefficients (Manichaikul et al. 2010) to identify both replicate samples and related individuals. Initially ANGSD is used to estimate genotype likelihoods (-doGlf 3) which are then leveraged by NgsRelate in a maximum likelihood framework to calculate R0, R1 and KING coefficients. Through the use of genotype likelihoods, NgsRelate takes the uncertainty of the genotypes into account, which is particularly important when working with low-coverage data, as inferring genotypes can be highly error prone and lead to an underestimation of heterozygous sites (Nielsen et al. 2011).

Pairs of samples with a KING-robust kinship coefficient > 0.125 were examined and the lower depth sample was discarded before further analysis. This included discarding the lower quality sample of any replicate pairs. For the later analyses using merged replicate blow samples, outliers were identified using the standard Tukey method applied within each replicate group. Specifically, for each level of the replicate factor, the first (Q1) and third (Q3) quartiles of the KING coefficient distribution were computed, and the interquartile range (IQR) was defined as $\text{IQR} = \text{Q3} - \text{Q1}$. A data point was classified as an outlier if it fell below $\text{Q1} - 1.5 \times \text{IQR}$ or above $\text{Q3} + 1.5 \times \text{IQR}$.

The eight replicate outlier pairs were then re-examined to see if a false identification had been made in the field. This included the careful assessment of drone-collected body imagery, because the scarring on the back and rostrum of the individual can be used to corroborate a correct matching ID (Figure 1B). Outliers were removed before a non-parametric Mann–Whitney *U* test with continuity correction was used to test the null hypothesis that replicate pairs have no higher relatedness than non-replicates, which would invalidate the method, with our expectation being that replicate pairs will have a significantly higher KING-robust kinship mean than non-replicate pairs. If a false ID was confirmed, the pair was no longer considered a replicate set and therefore not included in the merged dataset.

2.10 | Sex Assignment

To estimate the sex of sampled individuals and their replicates, we compared normalised read depths between the X chromosome and an autosome of similar length for all unmerged samples. Chromosome sizes in the *Megaptera novaeangliae* reference genome were manually inspected, and chromosome 6 (GenBank accession number CM085700.1) was selected as the closest in size to the X chromosome (GenBank accession number CM085716.1; chromosome 6 is 89.4% the size of chromosome X). We used SAMtools v.1.17 *idxstats* to extract per-chromosome read counts from aligned BAM files. The ratio of reads mapping to the X chromosome versus chromosome 6 was then computed for each individual. Based on expected genomic configurations, females (XX) were anticipated to exhibit a ratio near 1, and males (XY) a ratio near 0.5. We compared our sex-assignments to previously determined sexes based on field observations of known mothers, as well as genetic profiles from O'Mahony et al. (2024).

2.11 | Heterozygosity

Individual and population-level genomic diversity (heterozygosity) was estimated by realSFS implemented within ANGSD v. 0.940-2 (Korneliussen et al. 2014), using the default settings, to estimate the site frequency spectrum (SFS). This was done using saf files generated by ANGSD and using strict quality thresholds (-GL 2 -doSaf 1 -minMapQ 30 -minQ 30 -skipTriallelic 1 -uniqueOnly 1 -remove_bads 1 -only_proper_pairs 1) and for three different thresholds of minimum read depth at each considered site (minimum depth of 3×, 7× and 10×; -setMinDepth). We used the GATK genotype likelihood model (-GL 2) for SFS-based inference because it has been shown to outperform the SAMtools model when the average coverage of samples is low ($\leq 4\times$; Lou et al. 2021). We used the per-sample proportion of heterozygous loci from this one-dimensional SFS to estimate heterozygosity for each individual, following methods in Pečnerová et al. (2024).

To explore the effect of depth of coverage on within-sample heterozygosity estimates, we recalculated heterozygosity in depth bins of 3×, 4×, 5× and 6× to 200×. Sites $\geq 6\times$ were pooled into a single bin because per-depth site counts decline rapidly beyond ~5× and this pooled bin still comprised a substantial fraction of each sample's covered sites (mean = 6.23%), supporting a reasonably stable per-bin heterozygosity estimate. We ran a linear mixed model on the QC-filtered dataset ($n = 22$) to establish the effect of higher depth bins on heterozygosity estimates. Depth bins were treated as an ordinal fixed effect and sample as a random intercept. Based on these results, final humpback whale heterozygosity estimates were calculated using a minimum depth threshold of 3× and a maximum depth threshold of two-times each sample's mean depth of coverage (-setMaxDepth), thereby correcting for the upper tail of each sample's depth distribution (Figure S9), where collapsed repeats and paralogues generate artificial heterozygous calls. To further quantify the impact of sequencing depth on heterozygosity estimates, we

serially subsampled our two mid-coverage biopsy genomes (15× to 16×) to lower depths (0.25× to 15×, at 0.25×, 0.5× and 1× intervals), calculated per-site heterozygosity at each depth, and fit a linear regression model with sequencing depth as a continuous predictor.

To place the estimated heterozygosity of blow samples in the context of publicly available data, raw reads of available humpback whale genomes were pulled from the NCBI database (Table S4). The data were mapped to the same reference genome as the blow samples and filtered using the same pipelines. There was only one additional whole genome available stemming from the North Pacific (DNA Zoo, accession number SRR17854487).

To assess whether blow sampling an individual in the presence of other whales resulted in contamination, we compared heterozygosity estimates calculated with a minimum depth threshold of 3× between samples collected from solitary whales and samples collected when multiple whales were present (i.e., group size >1, where a group is considered to be whales within two body lengths of each other; Wray et al. 2021), on the basis that heterozygosity increases with increasing level of contamination (see section *Simulated sample contamination*). Differences in mean heterozygosity between groups were assessed using Welch's two-sample *t*-test, which does not assume equal variances between groups. The test was conducted with a directional alternative hypothesis that heterozygosity would be higher in samples of group size >1, as expected if contamination from additional individuals contributes to increased allelic diversity.

2.12 | Simulated Sample Contamination

Despite several precautions taken in the field to minimise the risk of sample contamination in the case of sampling individuals within a group (see section *Sample collection*), we recognise that intraspecific contamination might still occur in this context. Such within-species contamination can be difficult to detect and could result in ambiguous or incorrect genotype calls, particularly at low depth of coverage where each allele has only a small number of reads mapping to it (Jun et al. 2012). Such contamination might result in inflated heterozygosity and false identification of heteroplasmic mitogenome variants. To determine signals of contamination, we simulated sample contamination by randomly extracting a fraction of reads from one biopsy sample (subsampled to 2× to be most similar to blow samples) and introducing them into the remaining fraction of the other 2× biopsy subsample at varying proportions (1%–5%, 10%, 20%, 30%, 40% and 50% contamination). This was done using SAMtools *view* and *merge* commands. We calculated heterozygosity for these simulated contaminated samples following the same steps as above and used the increasing trend in heterozygosity with increasing contamination (Figure S10) to uphold the use of heterozygosity as a proxy for contamination in the *t*-test.

2.13 | Principal Component Analysis

We ran a principal component analysis (PCA) on the post-QC individuals using PCangsd v.1.36.3 (Meisner and Albrechtsen 2018) with genotype likelihoods (thereby accounting for missingness due to the low coverage of the blow samples). We did this first for the unmerged dataset ($n=22$) and again for the merged dataset ($n=20$), to determine expected population clustering of blow samples from the North Pacific separate from the North Atlantic (Figure 3C). For these analyses, we added the mid-coverage nuclear genomes from our two biopsied animals (and omitted their blow genomes), as well as publicly available mid- to high-coverage nuclear genome data from North Pacific and North Atlantic humpback whales (see Table S4 for accession numbers). These analyses had second-degree relatives removed (KING coefficient >0.125; discussed in section *Relatedness*). We did not expect to see any population structuring evident amongst the North Pacific individuals due to the assumption that all individuals stem from one interbreeding population and have been sampled within a relatively small geographical area, in comparison to the range of the North Pacific-wide population. We used ANGSD v.0.940-2 (Korneliussen et al. 2014) to generate genotype likelihoods using the GATK model (-GL 2) in beagle format (-doGlf 2). Strict quality filters were applied as described above, with additional filters including a low minor allele frequency (MAF) threshold of 5% and a likelihood ratio test for SNPs, retaining only sites where the probability of being polymorphic is high, that is, a *p*-value $\leq$ 1e-6 (-minMaf 0.05 -SNP_pval 1e-6). The minimum depth threshold was set to 3× (-set MinDepthInd 3) for both unmerged and merged runs.

2.14 | Runs of Homozygosity

To determine the efficacy of blow-derived genomes for estimation of inbreeding, we used ROHan (Renaud et al. 2019) to estimate the ROH present in the unmerged blow, biopsy and public genomes. We calculated the proportion of the genome in ROH and the mean ROH length per individual to assess patterns of homozygosity and potential inbreeding. We further counted the ROH present in each sample and categorised them according to their length (<1 Mb, 1–5 Mb and >5 Mb). For comparison between low and medium coverage genomes, we additionally estimated ROH in the biopsy samples which had been subsampled down to 2× and highlighted these in our summary figure (Figure 3E). This includes the whale BCX0860 from blow and biopsy at 2× and 15×, and whale BCX1596 from blow and biopsy at 2× and 16×. We used R v.4.4.0 (R Core Team 2015) and ggplot2 (Wickham 2016) for all visualisations.

3 | Results

3.1 | Samples, Sequencing Depth and Endogenous Content

We obtained WGS data from 58 blow samples representing 26 field-identified individuals with up to four replicate samples. Of these samples, 45 were collected in 2023 and 44 during the month of August (July = 7; September = 7). Further, 22 were collected during flights launched from land and 36 collected during flights launched from the bow of our research vessel (a 12 m aluminium skiff). Average whale group size during sampling events was 1.7 (min = 1; max = 5). Across the blow sample dataset, we got an average of 55.6 million read pairs, corresponding

to 111.2 million total reads, yielding an average endogenous content (i.e., percentage of reads mapping to the host species) of 83.6% (SD = 12.8%; min = 39.5%; max = 96.2%, Figure 2A). This left an average of 18,090,163 reads (16.4% of total reads) stemming from the microbiome of the respiratory tract of the sampled whale and from residue seawater, not analysed here.

The humpback whale low-coverage nuclear genomes obtained from the 58 (unmerged) blow samples had a mean depth of 2.3× (±2.4× SD; min = 0.2×; max = 3.5×; see Table S1). After the QC removal of three samples of less than 1× coverage, the blow samples had a mean depth of 2.45× (±2.3× SD; min = 1.03×; max = 3.5×). There was a highly significant positive correlation between a sample's depth of coverage and its endogenous content (Spearman's rank correlation rho = 0.7, $p < 10^{-16}$; Figure S1). Deeper sequencing of blow samples to 20 Gb (double the current 10 Gb) is predicted, per preseq, to yield a 1.47× to 1.98× increase in distinct reads (median = 1.81; mean = 1.77; $n = 55$ after QC removal of the three < 1× samples Figure S2). At current sequencing effort we cover on average 87% of the genome breadth (median = 89%; min = 66%; max = 91%); doubling effort would increase this to 94% on average (median = 96%; min = 73%; max = 97%) after QC removal of < 1× samples.

The blow samples stemmed from a mean of 2.6 breaths per sampling event (median = 3; min = 1; max = 6). The number of breaths collected was positively associated with depth of coverage. Linear regression indicated that each additional breath collected increased the expected sequencing depth by approximately 0.27× ($\beta = 0.272 \pm 0.070$ SE, $p < 0.001$). However, the number of breaths explained only 22% of variation in sequencing depth ($R^2 = 0.22$), indicating that additional factors also contribute to sequencing yield. Treating number of breaths as a categorical variable produced a similar conclusion, with significant differences in sequencing depth amongst breath collection groups ($F_{5,49} = 3.65$, $p = 0.007$), suggesting that the association was not driven solely by high-breath samples. Model assumptions were met based on residual diagnostics, including normally distributed residuals (Shapiro–Wilk $W = 0.990$, $p = 0.916$), and the estimated effect was unchanged in a robust regression sensitivity analysis ($\beta = 0.264$).

Merging of replicate blow samples from the filtered dataset resulted in genomic data with a mean depth at 5.6× and a variance of 4.2× (min = 2.7×; max = 11×; Table S1). For a single individual (BCX1484), we obtained a coverage of 11× based on four blow samples: two collected in 2022 and two in 2023. When the unmerged pre-QC blow sample dataset was mapped to a humpback reference mitogenome, the mean endogenous content was 0.04% (±0.012% SD) and the mean depth across the blow samples was 94× (±16.3× SD; min = 7.7×; max = 151.3×). We found a strong positive correlation between the depth of coverage of the nuclear genomes and the mitogenomes for the unmerged samples ($R^2 = 0.66$, $p = 7.5e\text{-}15$) and merged samples ($R^2 = 0.39$, $p = 0.0015$;

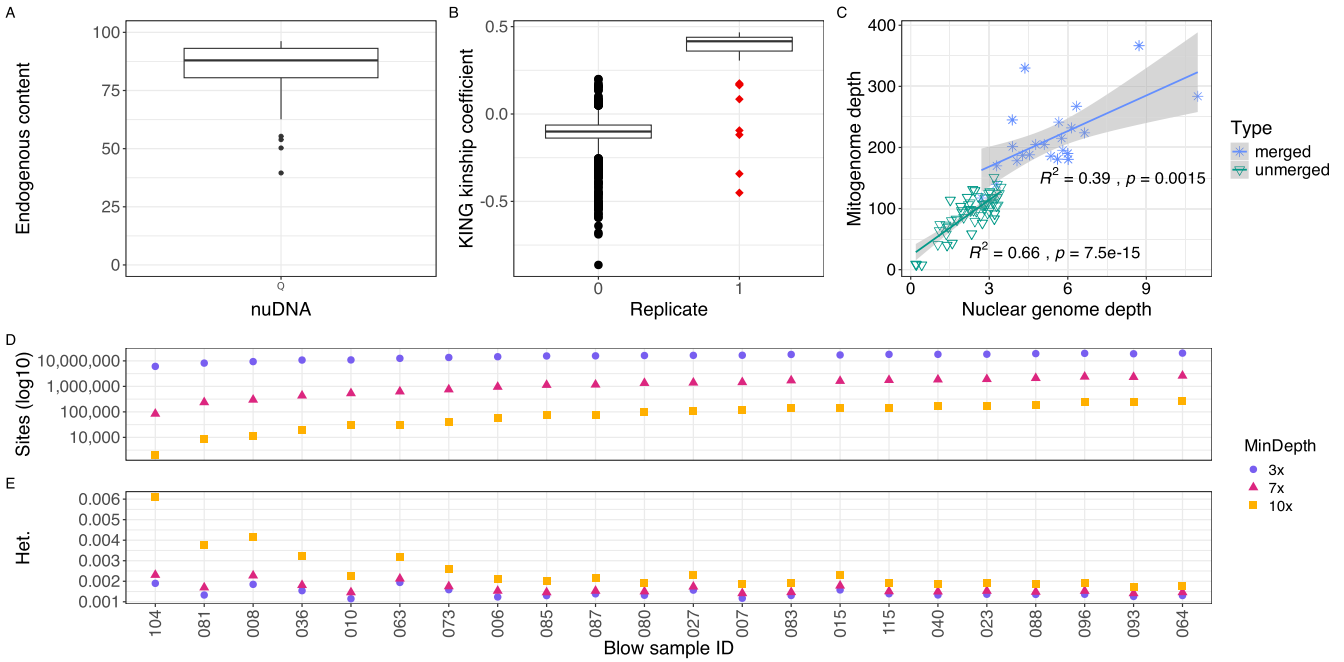


FIGURE 2 | (A) The endogenous content of unmerged blow samples for nuclear DNA (mean = 83.6% +/− 12.8%). (B) KING-robust kinship coefficients calculated without a filter for minimum depth, sorted on the *x*-axis according to field-assigned identification replicates (=1) and non-replicates (=0). Points highlighted as red diamonds were found to have incorrect field-assigned IDs and were removed during quality control. (C) Relationship between nuclear and mitochondrial sequencing depth of coverage across sample types. The scatter plot shows nuclear genome depth (*x*-axis) versus mitogenome depth (*y*-axis) for each sample, coloured and shaped by dataset type ('merged' and 'unmerged' blow samples). Linear regression lines are fitted separately for each type, with shaded areas representing 95% confidence intervals. Pearson's correlation coefficient (R^2) and associated *p*-values are shown for each regression. (D) The number of sites available for autosomal heterozygosity calculation, shown on a $\log_{10}$ scale, when a filter of minimum depth of 3×, 7× and 10× was applied. (E) Autosomal heterozygosity (Het.) of the final dataset of unmerged blow samples (worse quality replicates [i.e., lower depth of coverage] and highly related individuals removed; 22 samples remaining) using genotype likelihoods with varying per site minimum depth thresholds. Samples are ordered from left to right in an increasing number of sites available. Mean number of sites across samples: Min. depth 3× = 14,979,363; 7× = 1,246,257, 10× = 101,987. Colour blind-friendly palette from R package ggpubfigs (Steenwyk and Rokas 2021).

Figure 2C). The two biopsy samples had nuclear genome depths of 15× and 16× respectively and were subsequently serially down sampled to a minimum of 0.25× to evaluate how sequencing depth impacted the estimation of heterozygosity (Figure S8).

3.2 | Mitogenome Haplotypes and Heteroplasmy

When the mitogenome dataset was filtered to unique individuals only but retaining samples MN-b-089 and MN-b-001 (as these had unknown IDs), three previously described mtDNA control region haplotypes were present (A− = 48.3%; A+ = 27.6%; E2 = 10.3%), and we also found three new ('N') haplotypes (N1 = 6.9%; N2 = 3.4%; N3 = 3.4%). De novo mutations evident as heteroplasmy characterise these three previously unreported control region haplotypes and occur at nucleotide positions 15,640, 15,802 and 15,822 respectively (Table S2). Each of the new haplotypes were only observed in single individuals but importantly were consistent across replicate samples of those individuals. This, along with the high depth of coverage of the mitogenomes, drastically reduces the likelihood of this signal stemming from sequencing error or intraspecific contamination, although further investigation of these haplotypes is warranted. Suárez-Menéndez et al. (2023) also detected three new haplotypes stemming from heteroplasmy in their study of North Atlantic humpback whales, albeit at different nucleotide positions along the mitogenome. For comparison, Baker et al. (2013) found similar haplotype frequencies in northern BC of A− = 57.5%; A+ = 32.1%; E1 = 0.9%; E2 = 6.6%; F2 = 0.9% and total heteroplasmy = 1.9%, from 106 samples collected primarily by biopsy dart (and some sloughed skin).

3.3 | Sex Assignment and Relatedness

Sex of the sampled individuals and replicates was assigned by comparing the ratio of reads mapping to the X chromosome (135,787,239 bp) with those mapping to chromosome 6 (121,350,076 bp). Accounting for the difference in chromosome length, the expected read ratio is ~1.12 for females and ~0.56 for males. Ratio distributions were clearly separated between the sexes, with females showing a median ratio of 0.95 (range 0.82–1.00) and males a median ratio of 0.56 (range 0.53–0.67; Figure S3; Table S5), the latter closely matching theoretical expectations, while the former was somewhat lower, likely reflecting reduced mappability of the X chromosome (Webster et al. 2019). Nonetheless, this separation allowed unambiguous sex assignment for all individuals and replicates. Of the whales with known sex—either assigned by blow sample in O'Mahony et al. (2024) or in the field as observations of an adult mother closely associated with a calf—all sex assignments matched between methods (error rate = 0%). Further, all samples in the final post-QC dataset had matching sex assignment between replicates (error rate = 0%), supporting confident sex assignment using this method with samples of > 1× depth of coverage and passing other QC steps.

To estimate pairwise relatedness, over 6.5 million sites had genotype likelihoods assigned in preparation for the NgsRelate run (sites = 6,554,590). We re-evaluated the photo-ID and drone footage of any 'replicate' pairs with mismatching mtDNA haplotypes and low relatedness coefficients and discarded the mismatched sample from the merged dataset (Table 1 and red diamond-shaped outliers in Figure 2B). An overall error rate of misidentification in the field was 4 out of 56 samples (7%). Once misidentified pairs were reassigned as non-replicates, non-replicate pairs had a mean relatedness coefficient of −0.11, while replicate pairs had a mean relatedness of 0.41 (Figure S4), a highly statistically significant difference (Mann–Whitney U test $W = 0$, $p < 2.2e\text{-}16$).

3.4 | Population Genomic Inference

We demonstrated the applicability of our blow-derived low-coverage nuclear genome data for population genomic studies by inferring humpback whale population structure and genetic diversity using both the unmerged (lower coverage) nuclear genomes, merged (higher coverage) nuclear genomes and the mitogenomes, as well as the biopsy-derived genomes and published data (Figure 3A).

3.4.1 | Data Quality Control for Unmerged and Merged Blow Genome Datasets

Of the 58 blow samples sequenced, three samples were excluded because they had < 1× depth of coverage, 30 samples were removed as the lower depth replicates, two samples were excluded due to belonging to highly related pairs (KING > 0.125) and a further two samples were excluded because they fell above the 95th percentile relative error rate using the 'perfect sample approach' (Orlando et al. 2013). This left 22 unmerged blow samples for population genomic analyses (Figure S5A).

When replicate samples were compared for mtDNA control region haplotypes, three individuals field assigned as having multiple blow samples were found to have contradicting haplotype assignments, despite high mitogenome coverage, which alerted us to potential misidentifications in the field. Two of these individuals (BCX0427 and BCY0474) still had two correctly matched

TABLE 1 | The use of blow sampling generated nuclear and mitochondrial genomes, and specifically mtDNA control region haplotypes, to identify misidentified individuals, which is then corroborated using the 4K video footage captured by drone during the sampling flight (also see Figure 1B,C).

Sample	Field-ID	Haplotype	mtDNA depth (± SD)
MN-b-006	BCX0147	N1	119.7 (±22.6)
MN-b-070		N3	63.9 (±17.3)
MN-b-002	BCX0427	A+	139.6 (±35.1)
MN-b-089		E2	58.3 (±11.0)
MN-b-090		A+	127.8 (±15.5)
MN-b-001	BCY0474	A−	85.4 (±36.9)
MN-b-084		A+	69.7 (±14.2)
MN-b-085		A+	117.5 (±15.9)

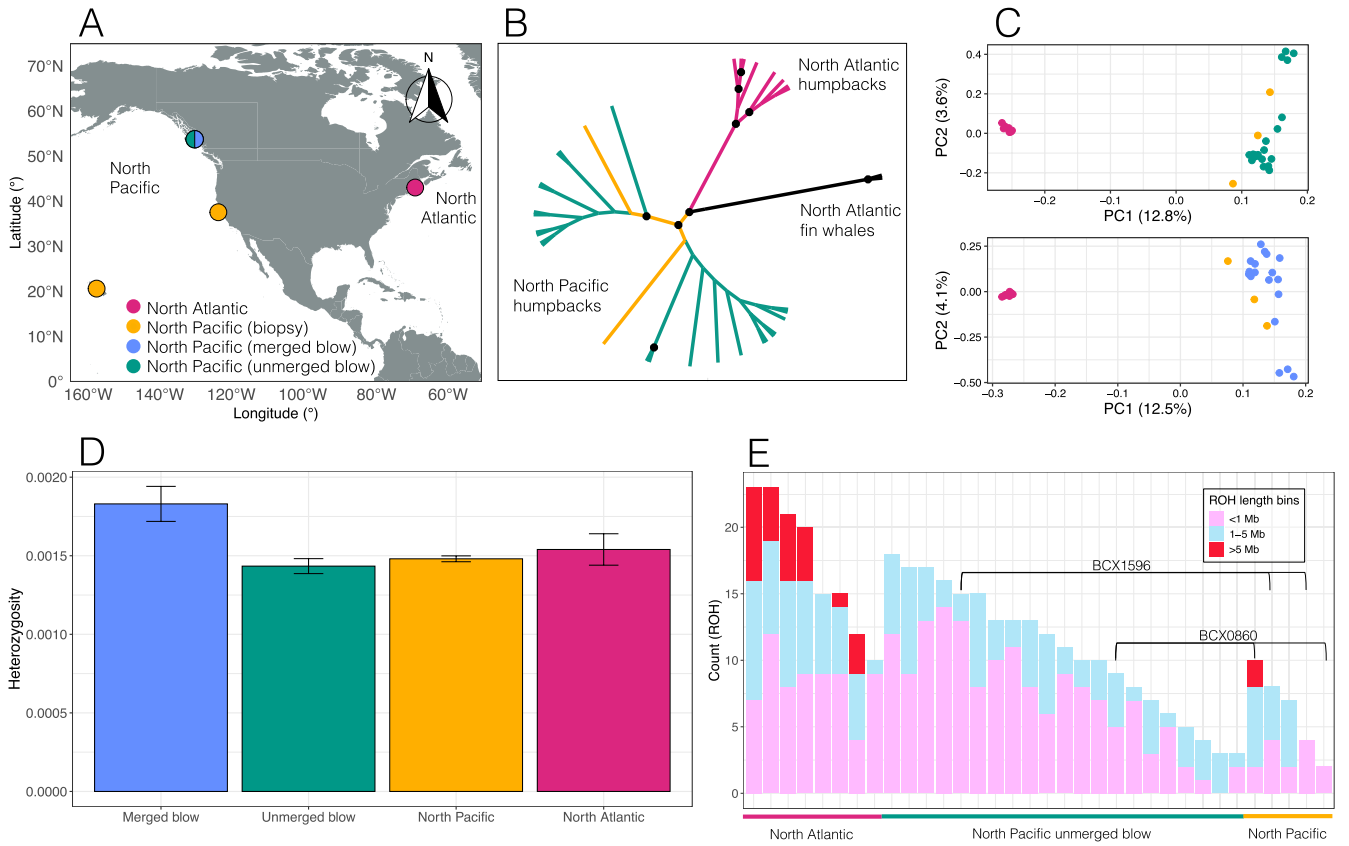


FIGURE 3 | Blow samples for population genomics and phylogenetics. (A) Map of North America showing sampling locations for blow samples collected in British Columbia, Canada (green and blue), for the two publicly available whole genomes from the North Pacific (yellow) and for the publicly available data from the North Atlantic (pink). (B) A cladogram constructed from consensus mitochondrial genomes, coloured according to panel A. Two North Atlantic fin whales are included as an outgroup and shown in black. Black circles indicate nodes with bootstrap support greater than 95% across 100 replicates. (C) A principal component analysis (PCA) using both the unmerged blow samples (green) and beneath this the merged blow samples (blue), as well as novel and publicly available genomic data based on biopsy samples from the North Pacific (yellow) and the North Atlantic (pink; see Table S4 for data sources). The PCAs were generated using genotype likelihoods at 5,936,486 SNPs and 6,348,584 SNPs for the unmerged and merged blow samples, respectively. (D) Autosomal heterozygosity of humpback whales estimated from blow samples (merged in blue and unmerged in green) compared with those generated by traditional biopsy methods. (E) Counts of the runs of homozygosity (ROH) per individual humpback whale (unmerged blow samples and biopsy samples), coloured according to their respective lengths (pale pink < 1 Mb; pale blue 1–5 Mb and red > 5 Mb). Samples are clustered according to their respective source populations/sample type. For comparison we have highlighted the blow samples with their associated biopsies, including both the subsampled 2× biopsies and full coverage biopsies for each whale (BCX0860 and BCX1596). The blow samples and 2× subsamples have higher counts of ROH than their respective higher depth biopsy.

blow samples which could be merged (Table 1). This was confirmed using pairwise relatedness estimates of the nuclear whole genomes, where pairs of samples with mismatching mtDNA haplotypes had lower than expected relatedness (Table S1).

For the QC of the merged dataset, three pairs of (merged) individuals were highly related (KING > 0.125; Table S3); and two individuals fell above the error rate threshold and thus these were removed from further analysis (Figure S5B). This left a total of 20 individuals in the final merged dataset. All samples flagged by the perfect sample approach also had elevated estimates of heterozygosity (Table S1).

3.4.2 | Phylogenomics and Population Structure

We found a high bootstrap support (100%) for the phylogenetic separation of the North Atlantic from the North Pacific humpback whales and a further subdivision of the North Pacific into two clusters (Figure 3B). One North Pacific cluster contained all A+, E2, N1 and N3 mtDNA haplotypes while the other cluster contained the A− and N2 haplotypes (Figure S6).

In line with the mitogenome data, we found a clear separation of genetic clusters according to ocean-basin by PCA, with no detectable population structure within the North Pacific (Figures 3C and S7).

3.4.3 | Genomic Diversity

We estimated autosomal heterozygosity for each individual humpback whale in the unmerged and merged datasets and placed these estimates in the context of heterozygosity estimates from the biopsy data and high coverage data using publicly available humpback whale genomes (Figure 3D). We found the

merged samples to have an elevated measure of heterozygosity when compared with unmerged blow samples and biopsy samples, suggesting the potential for further incorrect field identifications and as a consequence the merging of samples stemming from different individuals (as discussed above). We therefore recommend working with unmerged blow samples to avoid this potential error.

The unmerged post-QC blow sample mean heterozygosity across samples was 0.00141 ($n = 22$), which was slightly lower than the biopsy sample estimates of 0.00148 and 0.00154 stemming from the North Pacific and North Atlantic respectively. Estimates of heterozygosity presented here for the blow samples had a minimum depth threshold of 3× and a maximum depth threshold of twice the sample's mean depth of coverage (retaining a mean number of sites for genotype likelihood estimation = 13,618,924). We chose this threshold because when we varied the minimum depth threshold, which greatly affected the number of SNPs available for analysis, we found a minimum depth threshold of 3× to generate the closest estimates of heterozygosity compared with higher coverage data (Figure 2D). The mean number of sites across samples that were available for heterozygosity calculation at varying depth thresholds (without a maximum depth threshold applied) were as follows: min. depth 3× = 14,979,363; min. depth 7× = 1,246,257 and min. depth 10× = 101,987. Heterozygosity increased significantly with depth bin in the QC-filtered dataset at minimum depth of 3× ($n = 22$; Figure S9). A linear mixed model with depth bin as an ordinal fixed effect and sample as a random intercept showed a significant positive effect of depth bin on heterozygosity ($\beta = 4.8 \times 10^{-5}$ per bin, $SE = 5.0 \times 10^{-6}$; likelihood ratio test against the null model: $\chi^2(1) = 62.6$, $p < 0.001$). Consistent with this, heterozygosity in the ≥6× bin exceeded that in the 3× bin by 1.58×10^{-4} on average, an 11.4% increase relative to mean heterozygosity in the 3× bin (1.39×10^{-3}; paired t-test: $t_{21} = 8.85$, $p < 0.001$; Wilcoxon signed-rank test: $p < 0.001$). In comparison, heterozygosity estimates of subsampled biopsy-sourced genomes increased significantly with sequencing depth ($\beta = 1.08 \times 10^{-5}$ per 1× coverage across 0.25× to 15×, $R^2 = 0.44$, $p = 1.6 \times 10^{-6}$), indicating systematic underestimation at lower coverage (Figure S8). Heterozygosity estimates plateaued at approximately 4.5×, and given some replicate pair variability, we deduce that deeper sequencing of blow samples to at least this depth of coverage would result in more reliable data for estimation of heterozygosity.

When examining heterozygosity estimates of replicate pairs, some pairs showed unexpected variability. For example, for animal BCX1225, the samples MN-b-045 (2.3× depth; heterozygosity = 0.0012) and MN-b-103 (2.5× depth; heterozygosity = 0.0031) were collected 3 days apart. Sample MN-b-103 has higher heterozygosity yet was collected while the whale was alone and therefore cannot have an elevated heterozygosity estimate due to intraspecific contamination (Tables S1 and S5; Figure S10). Of the 22 unmerged post-QC samples, 15 were collected in single-whale encounters (i.e., no intraspecific contamination possible) and seven collected from individuals in a group size of two or more (but still only targeting the blow of one individual). Mean heterozygosity was similar between groups, with single-whale samples showing slightly higher heterozygosity (mean = 0.00143, SD = 0.0002) than the multi-whale samples (mean = 0.00136, SD = 0.00029). A Welch's two-sample t-test confirmed a lack of evidence that multi-whale samples exhibit elevated heterozygosity relative to single-whale samples ($t = -0.55$, $df = 8.81$, $p = 0.701$). This suggests that by following careful sampling protocols and a strict data QC pipeline, the presence of multiple whales during sampling events was not associated with increased heterozygosity estimates indicative of mixed-individual DNA contributions.

3.4.4 | Inbreeding

The unmerged blow samples had on average 0.70% (95% CI: 0%–7.9%) of their genomes in ROH, with an average ROH length of 1.5 Mb. The two biopsy-sourced genomes (BCX0860 and BCX1596) had 0.08% and 0.16% of their genomes in ROH respectively and an average ROH length of 1.0 Mb. For comparison, the single publicly available genome from the North Pacific (DNA Zoo, GenBank accession SRR17854487) had 0.56% (95% CI: 0.2%–0.7%) of its genome in ROH with an average ROH length of 1.8 Mb. The publicly available data from the North Atlantic had on average 3.7% (95% CI: 3.1%–4.5%) of their genomes in ROH, with an average ROH length of 4.7 Mb. We found lower counts of ROH in the biopsies (at full depth of 15× and 16×) compared with the individuals' blow samples (Figure 3E). The counts of ROH were also elevated in the 2× biopsy subsample, suggesting that higher coverage data are necessary to achieve accurate estimates of ROH and subsequent interpretation regarding inbreeding events in the population. Still, the overall low proportion of the genome in ROH estimated both from blow, biopsy and published genomes, together with the moderate average ROH length, indicates minimal recent inbreeding and suggests that homozygosity in these individuals is more likely the result of distant shared ancestry or historical demographic events such as population bottlenecks.

4 | Discussion

We have established that a complete humpback whale genome can be recovered from exhaled respiratory vapour ('blow'), marking a significant methodological advancement in non-invasive cetacean genomics. Prior to this, biopsies, strandings or historical specimens were required to obtain whole genome data—each with inherent limitations in availability, ethics or sample quality (Brown et al. 1994; Clapham and Mattila 1993; Keighley et al. 2021; Papastavrou and Ryan 2023). While in its infancy, blow sampling of cetaceans thus far has yielded data on hormonal profiles representative of life-history stages (Burgess et al. 2018); microbiota found within the respiratory tract (Apprill et al. 2017); and DNA of high enough quality to assign sex, mitochondrial haplotypes and enough microsatellite loci for individual identification (on average 1.7 loci across multiple species (Atkinson et al. 2021) and 7.5 loci for humpback whales; O'Mahony et al. 2024). Furthermore, we have generated a resource database of low-coverage whole genomes for North Pacific humpback whales which can be improved upon through genotype imputation using future high coverage data—something that is not possible with microsatellite data.

In our study, the endogenous content of the blow samples was 84% on average (SD = 12.8%; min = 39.5%; max = 96.2%), leaving an additional average of > 18 million reads per sample stemming from the microbiome of each whale's respiratory tract. This can further the study of cetacean holobionts, as from a single breath sample both host and microbiome genomes can be analysed. Paired with routine blow sampling for targeted study of hormonal profiles present within the breath (Burgess et al. 2018), blow becomes a valuable biological sample to assess individual to population-level health and resilience in the face of increasing anthropogenic pressures, including a rapidly changing global climate. An added benefit of blow samples over traditional biopsy sampling becomes especially apparent in contexts where biopsy sampling is not viable—for example, where strict non-invasive research protocols apply, where permitting and training pose a challenge or where whales actively evade approaching vessels due to human-wildlife conflict. Drone sampling eliminates the need for a boat if flying from land (37% of samples in this study), and/or avoids the need for a close vessel approach, which is required for both pole-based blow sampling and biopsy sampling (Hogg et al. 2009; Noren and Mocklin 2011). Given the size and mass of cetaceans, a close vessel approach poses risks to both the welfare of the researchers and the target species (Pirotta et al. 2017).

Several tests demonstrated the reliability and replicability of the blow-sourced genomic data, showing that the blow of free ranging whales can contain enough host DNA to generate on average 2.3× and up to 3.5× depth of coverage despite only a relatively low amount of data having been generated per sample (10 Gb per sample). Further validation was provided by the collection of replicate samples from the same individual humpback whales across two foraging seasons (25 individuals with 2–4 replicate samples each). We show that replicate pairs are genomic duplicates of each other with pairwise relatedness being significantly higher than those of non-replicate pairs ($p < 2.2e\text{-}16$), despite their low depth of coverage. Only 22% of the variability in sequencing depth was explained by the number of breaths collected per sampling event, although as expected, a higher number of breaths resulted in higher depth of coverage (each additional breath increased sequencing depth by 0.3×; $p < 0.001$). Additional factors contributing to the variability in sample quality likely include the whale's behaviour (and therefore the breath force), the exact height of the drone, whether the blow is 'wet' or 'dry' (O'Mahony et al. 2024) and the strength of the breeze. We eluted sample extracts to 150 μL in order to have sufficient material for quantification, sequencing and long-term storage. Future studies could consider eluting to lower volumes to increase DNA concentration, however the trade-off between higher concentration and total yield needs to be carefully considered if using a column-based extraction protocol as was used here.

Next-generation sequencing has revolutionised the fields of ecology and evolution in model and non-model organisms, especially the way that probabilistic genotype likelihood-based tools, such as ANGSD (Korneliussen et al. 2014) and ROHan (Renaud et al. 2019), give us the opportunity to carry out population genomics analyses on lower coverage data. In this study we have shown that despite sequencing the individual blow samples to relatively low coverage, a large number of SNPs are still available to carry out various population genomics tests—including PCAs which parse out the different populations, North Atlantic and North Pacific, as expected and measures of heterozygosity which correspond with estimates stemming from higher coverage biopsy-sourced genomes (although note that different library preparation and sequencing technologies were used for the biopsy samples). Finally, by rigorous testing and implementation of relatedness estimates in our study, we show that blow genomic data can be utilised for genetic mark-recapture studies (Palsbøll et al. 1997), as well as ground truthing of field photo-IDs of individual whales. Indeed, we had a small handful of instances in which estimated relatedness and mismatching mitochondrial haplotypes alerted us to putative false IDs made in the field.

Nonetheless, despite the successful demonstration of blow samples for genomic analyses, we recommend obtaining single blow samples with as many exhales as possible in the same sampling flight and sequencing these to a deeper coverage. In our simulations, increasing sequencing effort to obtain at least 4.5× depth of coverage may resolve variability in heterozygosity estimates (Figure S8 and Table S1) and estimates of ROH (Figure 3E). Considering the high endogenous content of these blow samples and the remaining availability of distinct reads for deeper sequencing (according to estimates of library complexity doubling sequencing effort to 20 Gb would have yielded a 1.5–2.0 times increase in unique reads, Figure S2), higher coverage blow genomes could have been obtained with deeper sequencing allowing for more in-depth population genomic analyses such as modelling of demographic histories and estimation of genetic load. This will also mitigate the risk of merging of samples from different individuals stemming from a false field-ID, which most likely explains the elevated heterozygosity visible in our merged blow dataset. For context, standard population genomic approaches based on WGS of tissue samples, 10 Gb, 20 Gb or 30 Gb of data would typically yield 4×, 8× and 12× coverage for a humpback genome at about 2.48 Gbp (Carminati et al. 2024).

The sampling approach presented here obviously comes with some challenges and limitations. Blow sampling of whales at sea is inherently weather dependent and sampling with a drone gets increasingly difficult as the wind picks up. However, this is also true for biopsy sampling, where wind, chop and swell all contribute to increasingly difficult sampling conditions. For blow sampling, the plume of exhalant gets flattened towards the sea surface in higher winds, requiring the pilot to navigate the drone closer to water. There is also variability in sample quality depending on several factors, including the pilots' skill and experience level, and the whale's behavioural state (e.g., resting humpback whales tend to have shallower breaths). Although generating whole genomes from the blow of other cetacean species has yet to be tested, we suggest blow sampling as a complementary genetic sampling strategy for large cetaceans that should be considered as an alternative to biopsy sampling when low-coverage WGS is sufficient to answer research questions. This is in line with the '3Rs' principle for animal research (replace, reduce, refine; Zemanova 2020), and is of value when a non-invasive approach is more suitable given the biological context, or when obtaining biopsies is not possible. We recognise that as a novel WGS tool, blow sampling will benefit from further testing and refinement, both in the field and in the laboratory. We also acknowledge that there will be situations where biopsy

sampling is necessary, for example, when deep-sequencing data are required or when archived tissue needs to be retained for additional DNA extraction and sequencing, or for parallel analyses of for example, stable isotopes, fatty acids or contaminant loads. Futhermore, applicability of blow sampling for genomics will likely vary across species depending on factors such as elusiveness, body size, swimming speed and surface time, for example with small, fast swimming odontocetes likely being more challenging to sample and containing less DNA per blow than humpback whales. Nevertheless, our demonstration that whale blow can yield complete genomes expands the possibilities of non-invasive sampling in marine biology. This approach enables population-scale genomics, longitudinal monitoring and evolutionary studies in cetaceans, without physical disturbance. Crucially, it offers a powerful tool for conservation, facilitating in situ assessments of population structure, genetic diversity and health, all key metrics for informed management in a rapidly changing ocean.

Author Contributions

Conceptualisation: É.N.O., M.L.M., O.E.G. and M.T.O. Methodology: É.N.O., M.L.M., O.E.G. and M.T.O. Blow sampling and associated lab work: É.N.O. Biopsy sampling: C.J.M. and S.J.T. Formal analysis: É.N.O. Validation: É.N.O., M.L.M., L.E.R., O.E.G. and M.T.O. Resources: É.N.O., J.W., C.J.M., S.J.T., O.E.G. and M.T.O. Data curation: É.N.O. Visualisation: É.N.O. and M.L.M. Writing – original draft: É.N.O. Writing – review and editing: É.N.O., M.L.M., E.M.K., J.W., C.J.M., S.J.T., L.E.R., O.E.G. and M.T.O. Supervision: M.L.M., L.E.R., O.E.G. and M.T.O. Project administration: É.N.O. Funding acquisition: É.N.O, J.W., O.E.G. and M.T.O.

Acknowledgements

We take this opportunity to thank the Gitga'at First Nation for their ongoing stewardship of their territory in which this work was carried out and for our continued collaboration to research and protect the whales foraging within the fjords. We would like to thank the field teams at the Fin Island Research Station and with Fisheries and Oceans Canada for their hard work and dedication to the whales of Gitga'at territory and beyond. A special thank you to Grace Baer, Nell O'Mahony, Florent Nicolas, Alma Foley, Abbey Lewis, Jemima Beddoe, Lisa Spaven, Elise Keppel, Jackie Hildering and Karina Dracott for their good humour and hard work in the field. We thank Chris Zadra from Ocean Alliance for his technical assistance, 3D-print creation of drone attachments and contributions to field work. We would like to thank Andreas Bak Pørksen for his contributions to lab work, as well as staff at the Globe Institute Molecular Biology Labs for their continued support. The computations were carried out on the HPC cluster 'Mjolnir' of the Globe Institute, University of Copenhagen, and we are grateful for its maintenance and the technical support provided by Bent Petersen.

Funding

The funding for this project was provided by the British Ecological Society's Small Research Grant (#SR23\1408), the Government of Canada Fisheries and Oceans Habitat Stewardship Program (Contribution Agreement 21-HSP-PAC-007), the Save Our Seas Foundation, the Donner Canadian Foundation, the Willow Grove Foundation, the World Wildlife Fund Canada, the School of Biology Research Committee University of St Andrews and the Carlsberg Foundation Semper Ardens Accelerate fellowship to MTO (#CF21-0425). Author ÉNO was supported by the Leslie & Charles Hilton Brown PhD Scholarship Fund (University of St Andrews School of Biology studentship), with training and travel grants provided by the Russell Trust Postgraduate Awards 2024 and the Genetics Society Training Grant 2024 (O'Mahony 2026).

Ethics Statement

This project was reviewed by the University of St Andrews' Animal Welfare and Ethics Review Body (Reference #BL16194).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Scripts used in the analyses are deposited on GitHub (https://github.com/eadinomahony/whale-blow-genomics-method) and sequencing data are available under the European Nucleotide Archive BioProject PRJEB93904.

References

Acevedo-Whitehouse, K., A. Rocha-Gosselin, and D. Gendron. 2010. "A Novel Non-Invasive Tool for Disease Surveillance of Free-Ranging Whales and Its Relevance to Conservation Programs." *Animal Conservation* 13, no. 2: 217–225.

Allendorf, F. W., P. A. Hohenlohe, and G. Luikart. 2010. "Genomics and the Future of Conservation Genetics." *Nature Reviews Genetics* 11, no. 10: 697–709.

Amos, W., H. Whitehead, M. J. Ferrari, D. A. Glockner-Ferrari, R. Payne, and J. Gordon. 1992. "Restrictable DNA From Sloughed Cetacean Skin; Its Potential for Use in Population Analysis." *Marine Mammal Science* 8, no. 3: 275–283.

Andrews, S. 2010. "FastQC: A Quality Control Tool for High Throughput Sequence Data." https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Apprill, A., C. A. Miller, M. J. Moore, J. W. Durban, H. Fearnbach, and L. G. Barrett-Lennard. 2017. "Extensive Core Microbiome in Drone-Captured Whale Blow Supports a Framework for Health Monitoring." *MSystems* 2, no. 5: 10–1128.

Árnason, Ú., F. Lammers, V. Kumar, M. A. Nilsson, and A. Janke. 2018. "Whole-Genome Sequencing of the Blue Whale and Other Rorquals Finds Signatures for Introgressive Gene Flow." *Science Advances* 4, no. 4: eaap9873.

Atkinson, S., A. Rogan, C. S. Baker, et al. 2021. "Genetic, Endocrine, and Microbiological Assessments of Blue, Humpback and Killer Whale Health Using Unoccupied Aerial Systems." *Wildlife Society Bulletin* 45, no. 4: 654–669.

Baker, C. S., D. Steel, J. Calambokidis, and E. Falcone. 2013. "Strong Maternal Fidelity and Natal Philopatry Shape Genetic Structure in North Pacific Humpback Whales." *Marine Ecology Progress Series* 494: 291–306. https://www.int-res.com/abstracts/meps/v494/p291-306/.

Borowska, E. I., Z. Nowak, C. van Elk, and M. Wahlberg. 2014. "Determining Genotypes From Blowhole Exhalation Samples of Harbour Porpoises (*Phocoena phocoena*)." *Aquatic Mammals* 40, no. 4: 407.

Brown, M. R., P. J. Corkeron, P. T. Hale, K. W. Schultz, and M. M. Bryden. 1994. "Behavioral Responses of East Australian Humpback Whales *Megaptera novaeangliae* to Biopsy Sampling." *Marine Mammal Science* 10, no. 4: 391–400.

Burgess, E. A., K. E. Hunt, S. D. Kraus, and R. M. Rolland. 2018. "Quantifying Hormones in Exhaled Breath for Physiological Assessment of Large Whales at Sea." *Scientific Reports* 8, no. 1: 10031.

Byrne, M. S., D. M. Peralta, E. A. Ibañez, and J. Nardelli M Túnez. 2021. "Non-Invasive Sampling Techniques Applied to Conservation Genetic Studies in Mammals." In *Molecular Ecology and Conservation Genetics of Neotropical Mammals*. Springer Nature.

Cammen, K. M., K. R. Andrews, E. L. Carroll, et al. 2016. "Genomic Methods Take the Plunge: Recent Advances in High-Throughput Sequencing of Marine Mammals." *Journal of Heredity* 107, no. 6: 481–495.

Carminati, M.-V., V. L. Gashi, R. Li, et al. 2024. "Novel *Megaptera novaeangliae* (Humpback Whale) Haplotype Chromosome-Level Reference Genome." *Scientific Data* 11, no. 1: 1113.

Carroll, E. L., M. W. Bruford, J. A. DeWoody, et al. 2018. "Genetic and Genomic Monitoring With Minimally Invasive Sampling Methods." *Evolutionary Applications* 11, no. 7: 1094–1119.

Centelleghe, C., L. Carraro, J. Gonzalvo, et al. 2021. "Correction: The Use of Unmanned Aerial Vehicles (UAVs) to Sample the Blow Microbiome of Small Cetaceans." *PLoS One* 16, no. 1: e0246177.

Chen, S., Y. Zhou, Y. Chen, and J. Gu. 2018. "Fastp: An Ultra-Fast All-In-One FASTQ Preprocessor." *Bioinformatics* 34, no. 17: i884–i890.

Christiansen, F., M. L. K. Nielsen, C. Charlton, L. Bejder, and P. T. Madsen. 2020. "Southern Right Whales Show no Behavioral Response to Low Noise Levels From a Nearby Unmanned Aerial Vehicle." *Marine Mammal Science* 36, no. 3: 953–963.

Clapham, P. J., and D. K. Mattila. 1993. "Reactions of Humpback Whales to Skin Biopsy Sampling on a West Indies Breeding Ground." *Marine Mammal Science* 9, no. 4: 382–391.

Costa, H., A. Rogan, C. Zadra, O. Larsen, A. H. Rikardsen, and C. Waugh. 2022. "Blowing in the Wind: Using a Consumer Drone for the Collection of Humpback Whale (*Megaptera novaeangliae*) Blow Samples During the Arctic Polar Nights." *Drones* 7, no. 1: 15.

Daley, T., and A. D. Smith. 2013. "Predicting the Molecular Complexity of Sequencing Libraries." *Nature Methods* 10, no. 4: 325–327.

de Flamingh, A., Y. Ishida, P. Pečnerová, et al. 2022. "Combining Methods for Non-Invasive Fecal DNA Enables Whole Genome and Metagenomic Analyses in Wildlife Biology." *Frontiers in Genetics* 13: 1021004.

Degollada, E., N. Amigó, S. O'Callaghan, M. Varola, K. Ruggero, and B. Tort. 2023. "A Novel Technique for Photo-Identification of the Fin Whale, *Balaenoptera physalus*, as Determined by Drone Aerial Images." *Drones* 7: 220. https://doi.org/10.3390/drones7030220.

Dudchenko, O., S. S. Batra, A. D. Omer, et al. 2017. "De Novo Assembly of the *Aedes aegypti* Genome Using Hi-C Yields Chromosome-Length Scaffolds." *Science* 356, no. 6333: 92–95.

Dudchenko, O., M. S. Shamim, S. S. Batra, et al. 2018. "The Juicebox Assembly Tools Module Facilitates De Novo Assembly of Mammalian Genomes With Chromosome-Length Scaffolds for Under $1000." Preprint, bioRxiv, January 28. https://doi.org/10.1101/254797.

Frère, C. H., E. Krzyszczyk, E. M. Patterson, S. Hunter, A. Ginsburg, and J. Mann. 2010. "Thar She Blows! A Novel Method for DNA Collection From Cetacean Blow." *PLoS One* 5, no. 8: e12299.

Fuentes-Pardo, A. P., and D. E. Ruzzante. 2017. "Whole-Genome Sequencing Approaches for Conservation Biology: Advantages, Limitations and Practical Recommendations." *Molecular Ecology* 26, no. 20: 5369–5406.

Funk, W. C., J. K. McKay, P. A. Hohenlohe, and F. W. Allendorf. 2012. "Harnessing Genomics for Delineating Conservation Units." *Trends in Ecology & Evolution* 27, no. 9: 489–496.

Hogg, C. J., T. L. Rogers, A. Shorter, K. Barton, P. J. O. Miller, and D. Nowacek. 2009. "Determination of Steroid Hormones in Whale Blow: It Is Possible." *Marine Mammal Science* 25, no. 3: 605–618.

Hohenlohe, P. A., W. C. Funk, and O. P. Rajora. 2021. "Population Genomics for Wildlife Conservation and Management." *Molecular Ecology* 30, no. 1: 62–82.

Jun, G., M. Flickinger, K. N. Hetrick, et al. 2012. "Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data." *American Journal of Human Genetics* 91, no. 5: 839–848.

Kashiwagi, T., E. A. Maxwell, A. D. Marshall, and A. B. Christensen. 2015. "Evaluating Manta Ray Mucus as an Alternative DNA Source for Population Genetics Study: Underwater-Sampling, Dry-Storage and PCR Success." *PeerJ* 3: e1188.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata. 2002. "MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform." *Nucleic Acids Research* 30, no. 14: 3059–3066.

Katoh, K., and D. Standley. 2013. "MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability." *Molecular Biology and Evolution* 30: 772–780.

Keighley, X., M. H. Bro-Jørgensen, H. Ahlgren, et al. 2021. "Predicting Sample Success for Large-Scale Ancient DNA Studies on Marine Mammals." *Molecular Ecology Resources* 21, no. 4: 1149–1166.

Korneliussen, T. S., A. Albrechtsen, and R. Nielsen. 2014. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* 15, no. 1: 356.

Korneliussen, T. S., and I. Moltke. 2015. "NgsRelate: A Software Tool for Estimating Pairwise Relatedness From Next-Generation Sequencing Data." *Bioinformatics* 31, no. 24: 4009–4011.

Li, H. 2013. "Aligning Sequence Reads, Clone Sequences and Assembly Contigs With BWA-MEM." Preprint, arXiv, March 16. https://doi.org/10.48550/arXiv.1303.3997.

Li, H. 2017. *seqtk: Toolkit for processing sequences in FASTA/Q formats*. Github. https://github.com/89/seqtk/issues/89.

Li, H., B. Handsaker, A. Wysoker, et al. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25, no. 16: 2078–2079.

Lieber, L., S. Berrow, E. Johnston, et al. 2013. "Mucus: Aiding Elasmobranch Conservation Through Non-Invasive Genetic Sampling." *Endangered Species Research* 21, no. 3: 215–222.

Lou, R. N., A. Jacobs, A. P. Wilder, and N. O. Therkildsen. 2021. "A Beginner's Guide to Low-Coverage Whole Genome Sequencing for Population Genomics." *Molecular Ecology* 30, no. 23: 5966–5993.

Manichaikul, A., J. C. Mychaleckyj, S. S. Rich, K. Daly, M. Sale, and W.-M. Chen. 2010. "Robust Relationship Inference in Genome-Wide Association Studies." *Bioinformatics* 26, no. 22: 2867–2873.

Meisner, J., and A. Albrechtsen. 2018. "Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data." *Genetics* 210, no. 2: 719–731.

Montana, L., T. T. Bringloe, A. Bourret, et al. 2024. "Reduced Representation and Whole-Genome Sequencing Approaches Highlight Beluga Whale Populations Associated to Eastern Canada Summer Aggregations." *Evolutionary Applications* 17, no. 12: 1–20.

Neveceralova, P., E. L. Carroll, D. Steel, et al. 2022. "Population Changes in a Whale Breeding Ground Revealed by Citizen Science Noninvasive Genetics." *Global Ecology and Conservation* 37: e02141.

Neveceralova, P., V. Soukup, J. Stundl, J. K. Stafford, and P. Hulva. 2025. "From Blow to DNA: Citizen Science Advancing Genetic Research on Southern Right Whales." *Conservation Genetics Resources* 17: 89–99.

Nguyen, L.-T., H. A. Schmidt, A. von Haeseler, and B. Q. Minh. 2015. "IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies." *Molecular Biology and Evolution* 32, no. 1: 268–274.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. "Genotype and SNP Calling From Next-Generation Sequencing Data." *Nature Reviews. Genetics* 12, no. 6: 443–451.

Noren, D. P., and J. A. Mocklin. 2011. "Review of Cetacean Biopsy Techniques: Factors Contributing to Successful Sample Collection and Physiological and Behavioral Impacts." *Marine Mammal Science* 28, no. 1: 154–199.

O'Mahony, É. N. 2026. "Behaviour and Population Structure of Baleen Whales in the Northeastern Pacific Ocean, Doctoral Thesis (PhD), University of St Andrews." https://doi.org/10.17630/sta/1644.

O'Mahony, É. N., A. L. Sremba, E. M. Keen, et al. 2024. "Collecting Baleen Whale Blow Samples by Drone: A Minimally Intrusive Tool for Conservation Genetics." *Molecular Ecology Resources* 24, no. 8: e13957.

Orlando, L., A. Ginolhac, G. Zhang, et al. 2013. "Recalibrating Equus Evolution Using the Genome Sequence of an Early Middle Pleistocene Horse." *Nature* 499, no. 7456: 74–78.

Palsbøll, P. J., J. Allen, M. Bérubé, et al. 1997. "Genetic Tagging of Humpback Whales." *Nature* 388, no. 6644: 767–769.

Papastavrou, V., and C. Ryan. 2023. "Ethical Standards for Research on Marine Mammals." *Research Ethics* 19, no. 4: 390–408.

Pečnerová, P., E. Lord, G. Garcia-Erill, et al. 2024. "Population Genomics of the Muskox' Resilience in the Near Absence of Genetic Variation." *Molecular Ecology* 33, no. 2: e17205.

Pirotta, V., A. Smith, M. Ostrowski, et al. 2017. "An Economical Custom-Built Drone for Assessing Whale Health." *Frontiers in Marine Science* 4: 425.

Port, J. A., J. L. O'Donnell, O. C. Romero-Maraccini, et al. 2016. "Assessing Vertebrate Biodiversity in a Kelp Forest Ecosystem Using Environmental DNA." *Molecular Ecology* 25, no. 2: 527–541.

Quinlan, A. R., and I. M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics (Oxford, England)* 26, no. 6: 841–842.

R Core Team. 2015. "R: A Language and Environment for Statistical Computing, (Version 1.1.383)." https://www.R-project.org/.

Renaud, G., K. Hanghøj, T. S. Korneliussen, E. Willerslev, and L. Orlando. 2019. "Joint Estimates of Heterozygosity and Runs of Homozygosity for Modern and Ancient Samples." *Genetics* 212, no. 3: 587–614.

Richard, J. T., K. Schultz, C. Goertz, R. Hobbs, T. A. Romano, and B. L. Sartini. 2017. "Assessing the Quantity and Downstream Performance of DNA Isolated From Beluga (*Delphinapterus leucas*) Blow Samples." *Aquatic Mammals* 43, no. 4: 398–408.

Robinson, C. V., and H. K. Nuuttila. 2020. "Don't Hold Your Breath: Limited DNA Capture Using Non-Invasive Blow Sampling for Small Cetaceans." *Aquatic Mammals* 46, no. 1: 32–41.

Smit, A. F. A., R. Hubley, and P. Green. 2013. "RepeatMasker Open-4.0." http://www.repeatmasker.org.

Steenwyk, J. L., and A. Rokas. 2021. "ggpubfigs: Colorblind-Friendly Color Palettes and ggplot2 Graphic System Extensions for Publication-Quality Scientific Figures." *Microbiology Resource Announcements* 10, no. 44. https://doi.org/10.1128/mra.00871-21.

Suárez-Menéndez, M., M. Bérubé, F. Furni, et al. 2023. "Wild Pedigrees Inform Mutation Rates and Historic Abundance in Baleen Whales." *Science* 381, no. 6661: 990–995.

Székely, D., N. L. Corfixen, L. L. Mørch, et al. 2021. "Environmental DNA Captures the Genetic Diversity of Bowhead Whales (*Balaena mysticetus*) in West Greenland." *Environmental DNA* 3, no. 1: 248–260.

Tokarska, M., T. Marshall, R. Kowalczyk, et al. 2009. "Effectiveness of Microsatellite and SNP Markers for Parentage and Identity Analysis in Species With Low Genetic Diversity: The Case of European Bison." *Heredity* 103, no. 4: 326–332.

Tollis, M., J. Robbins, A. E. Webb, et al. 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, no. 8: 1746–1763.

Valverde, J. R., R. Marco, and R. Garesse. 1994. "A Conserved Heptamer Motif for Ribosomal RNA Transcription Termination in Animal Mitochondria." *Proceedings of the National Academy of Sciences of the United States of America* 91, no. 12: 5368–5371.

Visser, F., V. J. Merten, T. Bayer, et al. 2021. "Deep-Sea Predator Niche Segregation Revealed by Combined Cetacean Biologging and eDNA Analysis of Cephalopod Prey." *Science Advances* 7, no. 14: eabf5908.

Von Duyke, A. L., J. A. Crawford, L. Quakenbush, J. R. Adams, and L. P. Waits. 2023. "Determination of Polar Bear (*Ursus maritimus*) Individual Genotype and Sex Based on DNA Extracted From Paw-Prints in Snow." *Frontiers in Conservation Science* 4: 1166279. https://doi.org/10.3389/fcosc.2023.1166279.

Waits, L. P., and D. Paetkau. 2005. "Noninvasive Genetic Sampling Tools for Wildlife Biologists: A Review of Applications and Recommendations for Accurate Data Collection." *Journal of Wildlife Management* 69, no. 4: 1419–1433.

Webster, T. H., M. Couse, B. M. Grande, et al. 2019. "Identifying, Understanding, and Correcting Technical Artifacts on the Sex Chromosomes in Next-Generation Sequencing Data." *GigaScience* 8, no. 7: p.giz074.

Weinrich, M. T., R. H. Lambertson, C. R. Belt, and M. R. Schilling. 1992. "Behavioral Reactions of Humpback Whales *Megaptera novaeangliae* to Biopsy Procedures." *Fishery Bulletin* 90, no. 3: 588–598.

Wickham, H. 2016. "ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org/.

Wray, J., E. Keen, and É. N. O'Mahony. 2021. "Social Survival: Humpback Whales (*Megaptera novaeangliae*) Use Social Structure to Partition Ecological Niches Within Proposed Critical Habitat." *PLoS One* 16, no. 6: e0245409.

Zemanova, M. A. 2019. "Poor Implementation of Non-Invasive Sampling in Wildlife Genetics Studies." *Rethinking Ecology* 4: 119–132.

Zemanova, M. A. 2020. "Towards More Compassionate Wildlife Research Through the 3Rs Principles: Moving From Invasive to Non-Invasive Methods." *Wildlife Biology* 2020: 1–17.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. **Figure S1:** Scatterplot of unmerged blow sample depths of coverage (*x*-axis) versus each sample's endogenous content (*y*-axis) with associated Pearson correlation coefficient and *p*-value, demonstrating a significant positive correlation between the two variables. **Figure S2:** Blow sample library complexity extrapolation estimated using preseq v.3.1.2 and all 58 unmerged blow samples. Each sample's extrapolated curve is truncated at five-times its own depth. The dashed vertical line marks the current median depth, the dotted vertical line marks doubling this, while the dotted slope indicates a hypothetical perfect 1:1 relationship between total read pairs and distinct read pairs. The three worst samples and three best samples are flagged as examples with their respective overall depth of coverage. The three lowest complexity samples are the only three samples with depth of coverage $<1\times$ (and are therefore removed during QC filtering). **Figure S3:** The distribution of sample ratios of number of reads mapping to the X chromosome relative to chromosome six, where a ratio approaching 1 is a female (XX sex chromosomes), while a ratio approaching 0.5 will be male (XY sex chromosomes). **Figure S4:** KING-robust kinship coefficients for pairs of blow samples filtered according to resolved false field identifications and sorted according to field replicates (1) and non-replicates (0). This figure represents the post-quality control dataset. **Figure S5:** Error rates (%) calculated according to the 'perfect sample approach' in ANGSD. The red dashed lines represent each respective 95th percentile cutoff, above which threshold samples were considered too error prone and discarded from further analysis. (A)

The unmerged blow samples (*x*-axis) that passed prior quality control (QC) steps, resulting in the removal of samples MN-b-103 and MN-b-127. (B) The merged samples (*x*-axis) that passed prior QC steps, resulting in the removal of BCX0195 and CSY0149. Samples as blue squares are the biopsy-sourced genomes from our study, while the rest are blow sample-sourced. **Figure S6:** Mitogenome maximum likelihood phylogenetic tree with 100 replicate bootstraps, constructed using the GTR substitution model in IQ-TREE[3]. Samples include the unmerged post-QC blow samples (sample codes MN-b-***) which have their associated control region haplotype noted in brackets, the North Pacific humpback whale reference mitogenome, the DNA Zoo humpback whale reference mitogenome and eight North Atlantic humpback whale mitogenomes (GenBank accession numbers in brackets). Two North Atlantic fin whale mitogenomes were used as an outgroup to root the tree (NC_001321[4] and OZ239533 (Darwin Tree of Life submission, https://portal.darwintreeoflife.org)). **Figure S7:** A principal component analysis (PCA) using the unmerged blow samples that passed quality control (green) with the biopsy-sourced whole genomes from the North Pacific (yellow) and publicly available whole genomes from the North Atlantic (pink). **Figure S8:** Genome-wide heterozygosity of two humpback whale biopsy samples, subsampled to varying depths of coverage from 15× to 0.25×, demonstrating a decrease in estimated heterozygosity at lower coverages. The black vertical lines mark the mean heterozygosity of the sequenced blow samples analysed in this study. **Figure S9:** Assessment of the impact of higher depth of coverage sites on the estimation of heterozygosity ('Het') for unmerged post-QC blow samples (panels A and B) and merged post-QC blow samples (panels C and D). The effect of a maximum depth threshold (twice each sample's mean depth of coverage) applied to estimate of heterozygosity with a minimum depth threshold of 3× is presented in panels A and C, while heterozygosity estimates for varying depths bins (3×, 4×, 5× and 6× to 200×) are presented in panels C and D. **Figure S10:** Genome-wide heterozygosity of simulated contamination, generated from subsampled 2× biopsies of whales BCX0860 and BCX1596. The 2× subsamples were used to most closely represent the blow sample data analysed in this study. **Table S1:** Blow samples stemming from multiple samples of the same individuals. Reported heterozygosity estimates were calculated using a minimum depth threshold of 3× and a maximum depth threshold of twice each sample's mean depth of coverage. Whale IDs and blow sample IDs in red with an asterisk were flagged during quality control (QC): either too low depth of coverage, found to be mismatching or fell outside the error rate threshold set by the perfect sample approach and were removed during QC. Depth of coverage (± one standard deviation, SD) are reported for unmerged and resolved (i.e., only containing correct IDs) merged samples. Mitochondrial (mtDNA) control region haplotypes are reported according to the haplotypes reported in Baker et al. (2013)[5], with novel haplotypes from heteroplasmic sites simply named using the letter N for 'new': 'N1', 'N2' and 'N3'. **Table S2:** Potential occurrence of heteroplasmy in samples differing from known North Pacific mtDNA control region D-loop haplotypes. Even samples of low mitogenome coverage (MN-b-013 and MN-b-117) show heteroplasmy at the correct site and are in agreement with their higher coverage replicate samples. Samples with an asterisk were later determined to have a false field-ID, due to the discrepancies in heteroplasmic sites and KING-robust kinship coefficients. **Table S3:** Pairs of samples flagged as highly related in the NgsRelate run (KING-robust kinship coefficient > 0.125) using the merged blow samples, biopsies and publicly available data. Samples of higher coverage were retained for downstream analyses (and those removed highlighted with an asterisk). **Table S4:** Publicly available nuclear and mitochondrial genome resequencing data used for ground truthing the blow sample generated whole genomes in this study. Samples chosen from Suárez-Menéndez et al. (2023)[6] were from the parental generation, to avoid related dyads. **Table S5:** Blow sample overview, including European Nucleotide Archive (ENA) accession numbers (Project accession PRJEB93904), whale field identifications (ID), sex assignment (F = female; M = male), ratio of reads mapping to chromosome X to autosome 6 (X to 6 ratio), mitochondrial (mtDNA) control region haplotype, number of breaths sampled, whale group size, endogenous content, Qubit concentration and sequencing depth of coverage for nuclear (nuDNA) and mitochondrial (mtDNA) genomes.