ISCI, Volume *15*

Supplemental Information

Narwhal Genome Reveals Long-Term Low Genetic Diversity despite Current Large Abundance Size

Michael V. Westbury, Bent Petersen, Eva Garde, Mads Peter Heide-Jørgensen, and Eline D. Lorenzen

Supplemental information

Supplemental tables:

Supplemental table S1: Illumina and cross-mate libraries used for the assembly, Related to Results and discussion section Narwhal genome assembly.

Library	Approximate insert size	# of reads
Short insert 01	-	384,563,392
Short insert 02	-	379,311,470
Short insert 03	-	381,357,017
Mate-pair	3kb	152,848,424
Mate-pair	5kb	108,544,278
Mate-pair	10kb	100,789,266
Cross-mate	500bp	48,445,804
Cross-mate	1kb	45,509,780
Cross-mate	1.5kb	44,383,867
Cross-mate	2kb	43,559,749
Cross-mate	3kb	42,428,532
Cross-mate	4kb	41,623,118
Cross-mate	5kb	41,082,099
Cross-mate	8kb	40,265,781
Cross-mate	10kb	40,091,817
Cross-mate	15kb	39,771,013
Cross-mate	20kb	39,635,226

Supplemental table S2: BUSCO scores of the assembled narwhal genome when using the BUSCOv3 mammal dataset, Related to Results and discussion section Narwhal genome assembly.

Category	Number of BUSCO	Percentage
Complete	3819	93.00%
Complete and single copy	3772	91.90%
Complete and duplicated	47	1.10%
Fragmented	140	3.40%
Missing	145	3.60%
Total searched	4104	

Supplemental table S3: Narwhal genome repeat profile, Related to Results and discussion section Narwhal genome assembly.

Repeat type	De novo repeats (%)	Model based repeats (%)	Total (%)
Total	34.87	3.03	37.90
SINEs	6.26	0.22	6.48
LINEs	20.43	1.63	22.06
LTR elements	4.87	0.70	5.57
DNA elements	3.17	0.40	3.57
Unclassified	0.05	0.08	0.13
Small RNA	3.05	0.21	3.26
Satellites	0.07	0.00	0.07

Supplemental table S4: Standard deviation of heterozygosity in 500kb windows across the autosomes of the narwhal and other endemic Arctic marine mammals, Related to Figure 2.

Species	Standard deviation
Narwhal	0.0000642673
Beluga	0.000114535
Bowhead	0.000166954
Walrus	0.000191573
Polar bear	0.000272893

Supplemental table S5: Distribution of heterozygosity in different regions across the narwhal genome, Related to Figure 3.

Feature/location	Mean heterozygosity	Standard deviation
Autosomes	0.000138	0.00000333
Exons	0.000127	0.00000913
Exons and introns	0.000133	0.00000480
10kb away	0.000132	0.00000410
20kb away	0.000136	0.00000500
50kb away	0.000141	0.00000721

Supplemental table S6: Distribution of heterozygosity in different regions across the beluga genome, Related to Figure 3.

Feature/location	Mean heterozygosity	Standard deviation
Autosomes	0.000289	5.31e-06
Exons	0.000221	1.60e-05
Exons and introns	0.000277	6.21e-06
10kb away	0.000284	6.34e-06
20kb away	0.000285	7.36e-06
50kb away	0.000294	7.81e-06

Supplementary table S7: Unpaired two sample t-test to test for significant differences in heterozygosity between different genomic regions in the narwhal, Related to Figure 3.

Regions compared	t-score	p-value
Autosomes vs. exons	11.389	< 2.2e-16
Autosomes vs. genes	7.6497	8.59e-13
Autosomes vs. 10kb away	10.466	< 2.2e-16
Autosomes vs. 20kb away	3.013	0.002924
Autosomes vs. 50kb away	-4.1801	4.37e-05
Exons vs. genes	-6.3992	1.11e-09
Genes vs. 10kb away	1.6733	0.09585
Genes vs. 20kb away	-3.8386	0.0001664

Supplementary table S8: Unpaired two sample t-test to test for significant differences in heterozygosity between different genomic regions in the beluga, Related to Figure 3.

Regions compared	t-score	p-value
Autosomes vs. exons	42.691	2.20e-16
Autosomes vs. genes	14.707	2.20e-16
Autosomes vs. 10kb away	6.9906	4.09e-11
Autosomes vs. 20kb away	5.6237	6.30e-08
Autosomes vs. 50kb away	-7.4233	3.31e-12
Exons vs. genes	-42.691	2.20e-16
Genes vs. 10kb away	-34.921	2.20e-16
Genes vs. 20kb away	-38.454	2.20e-16

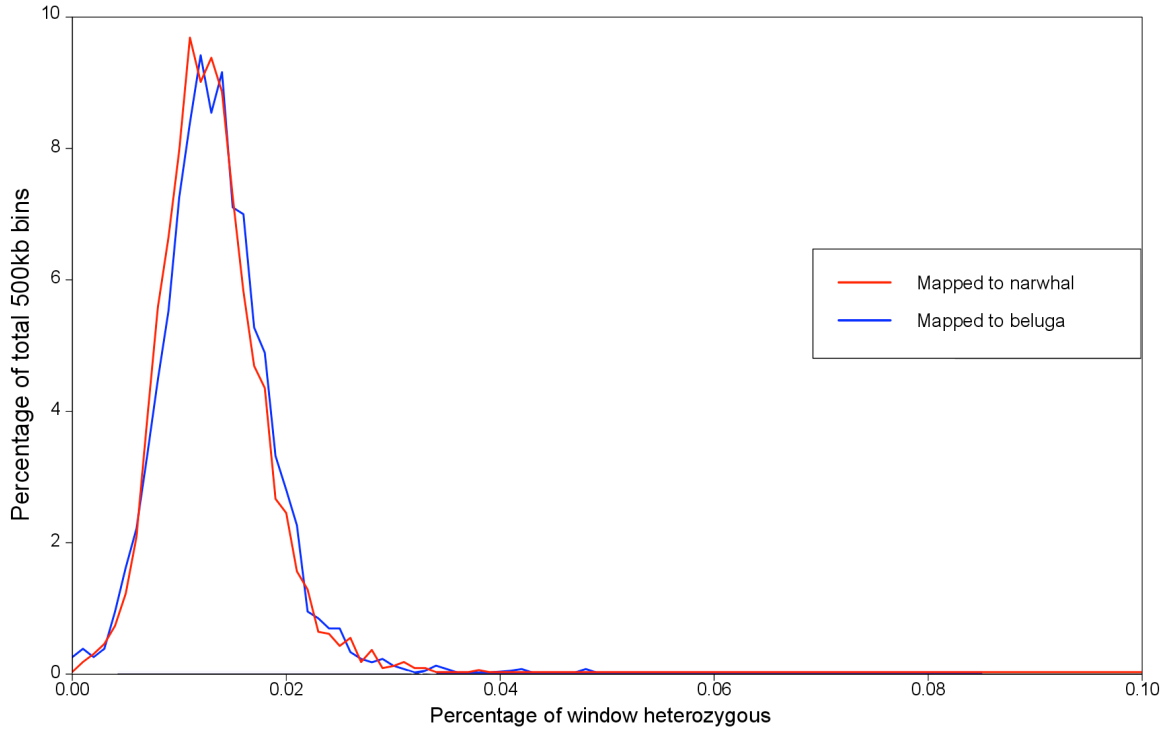
Supplemental table S9: Pairwise distances (PWD) and species divergence times used to calculate mutation rates and the resultant mutation rates per year, Related to Figure 4.

	PWD	Divergence time (Ma)	u/year
Bowhead-right whale	0.0067	4.38	0.0000000007676940639
Beluga-Narwhal	0.0057	5.5	0.0000000005181818182
Seal-Walrus	0.023	18	0.0000000006259166667

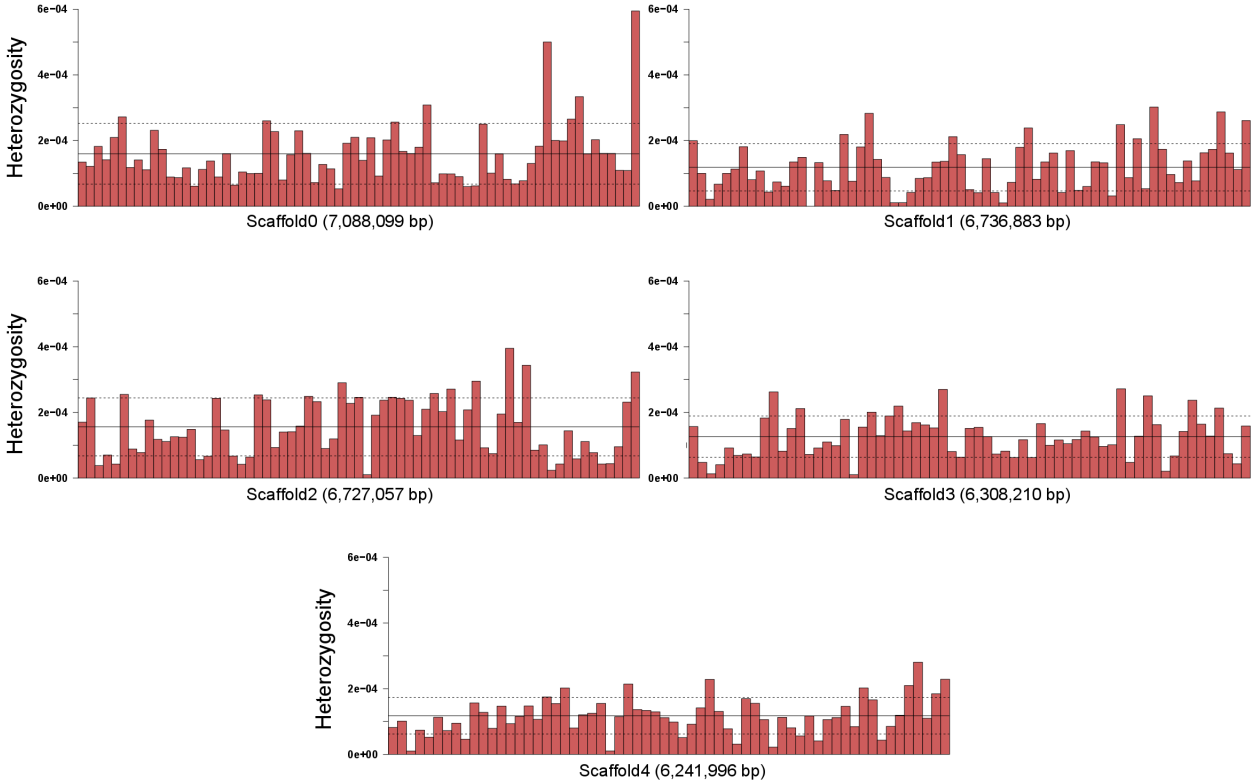
Supplemental table S10: Mutation rates and generation times used for plotting PSMC, Related to Figure 4.

Common name	Genus	Species	Generation time	u/generation
Narwhal	*Monodon*	*monoceros*	30	1.56e-08
Beluga	*Delphinapterus*	*leucas*	32	1.65e-08
Polar bear	*Ursus*	*maritimus*	11.2	1.83e-08
Walrus	*Odobenus*	*rosmarus*	15	9.40e-09
Bowhead whale	*Balaena*	*mysticetus*	35	2.69e-08

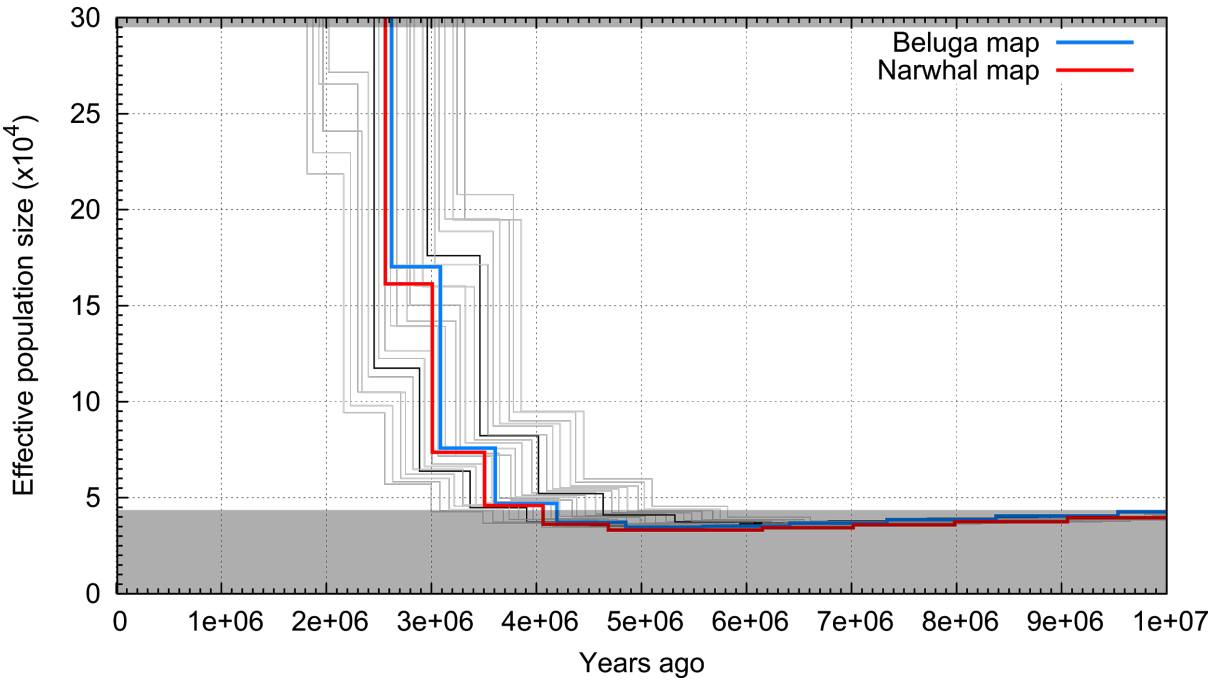
Supplemental figures:



Supplemental figure S1: Sliding window heterozygosity in the narwhal when using different mapping reference genomes, Related to Figure 2. Red shows when mapped to the narwhal. Blue shows when mapped to the beluga.



Supplemental figure S2: Heterozygosity in 100kb non-overlapping sliding windows across the five largest scaffolds from the narwhal assembly, Related to Figure 2. Black line represents the mean heterozygosity value for said scaffold and the dotted line represents one standard deviation above and below the mean heterozygosity value.



Supplemental figure S3: hPSMC plot between the beluga and the narwhal and simulations of various different divergences, Related to Results and discussion section Comparative history of the narwhal and beluga. Greyed out regions represent 1.5x and 10x the pre divergence effective population size, grey lines represent the simulated data, black line represents the simulations closest to the real data without overlapping it, blue line represents the hPSMC result when both the narwhal and beluga were mapped to the beluga reference genome, red line represents the hPSMC result when both the narwhal and the beluga were mapped to the narwhal reference genome.

Transparent methods:

Sample information

The narwhal individual was sampled in Uummannaq/West Greenland in 1993, and originated from the Somerset Island stock. It was collected from the Greenland Institute of Natural Resources under the general permit for biological sampling of the Inuit from the Greenland Government. The sample was exported to Denmark under CITES permit number 15GL1003549. The Somerset Island stock is one of the largest narwhal stocks with current population levels being estimated at ~50,000 individuals (NAMMCO 2018).

Genome assembly

Whole genomic DNA was extracted from a frozen liver tissue from a single narwhal individual using the QIAGEN DNeasy blood and tissue kit following the manufacturer's protocol with slight modifications (2x volume of reagents (except AW1 and AW2)). Extracts were built into three short insert Illumina sequencing libraries and three mate-paired Illumina sequencing libraries (~3kb ~5kb, ~10kb) by the UC Davis genome center (http://genomecenter.ucdavis.edu/). Libraries were sequenced at the UC Davis genome center on an Illumina HiSeq platform. Additionally, we constructed cross-species 100bp mate paired reads of insert sizes between 500bp and 20kb (Table S1) utilising the repeat masked beluga genome (Genbank: GCA_002288925.2) (Jones et al. 2017) and the software Cross-Species Scaffolding (Grau et al. 2018). We removed adapter sequences from the short insert and mate paired libraries using skewer (Jiang et al. 2014) and removed PCR duplicates with prinseq (Schmieder & Edwards 2011). We performed an error correction step using a kmer size of 31 in tadpole from the bbtools toolsuite (Bushnell 2014). We constructed a *de novo* assembly using these error corrected reads, the three mate paired libraries and the cross species mate paired libraries using SOAPdenovo2 (Luo et al. 2012) and specified a kmer size of 51. The short insert libraries were used in both the contig construction and scaffolding steps while the mate paired libraries were only used in the scaffolding step. We removed all contigs shorter than 1000bp from the final assembly. We performed gap closing on the assembly with Sealer (Paulino et al. 2015), utilising various kmer sizes (50, 60, 70, 80, 90, 100) and the error corrected short insert library reads. The assembly continuity was assessed using quast v4.5 (Gurevich et al. 2013) and gene content was assessed using BUSCO v3 (Waterhouse et al. 2017) and the mammalian BUSCO gene set database.

Repeatmasking and annotation

Repeats and low complexity DNA sequences were masked in the genome prior to gene annotation using RepeatMasker version open-4.0.7 (Smit et al. 2013-2015) using the species repeat database 'narwhal' with RepBase database version 20170127. Remaining specific repetitive elements were predicted *de novo* using RepeatModeler version 1.0.11 (Smit & Hubley 2008-2015) on the masked genome. Subsequently, a second round of RepeatMasker was run with the model generated from RepeatModeler as custom library input on the previously masked genome.

Genome annotation was performed using the genome annotation pipeline MAKER2 version 2.31.9 (Holt & Yandell 2011) with ab-initio and homology-based gene predictions. Protein sequences from killer whale (*Orcinus orca*), beluga whale (*Delphinapterus leucas*), cattle (*Bos taurus)*, dog, (*Canis lupus familiaris*), humans (*Homo sapiens*), minke whale (*Balaenoptera acutorostrata*) and the finless porpoise (*Neophocaena asiaeorientalis*) were used for homology-based gene prediction. As no training gene models were available for narwhals, we used CEGMA (Parra et al. 2007; Parra et al. 2009) to train the ab-initio gene predictor SNAP (Korf 2004), rather than using the de-novo gene predictor in Augustus

(Stanke & Waack 2003). MAKER2 was run with "model_org=simple, softmask=1, augustus_species=human" and the "snaphmm" parameter was set to the HMM generated in the manual training of SNAP. As EST evidence we used a published transcriptome skin sample of a beluga whale (Genbank: PRJNA414234).

Heterozygosity estimates

We estimated autosomal heterozygosity from our narwhal genome and four endemic Arctic marine mammals. We downloaded the assembled genomes and raw Illumina reads from the beluga (*Delphinapterus leucas* Genbank: GCA_002288925.2), bowhead whale (*Balaena mysticetus*) (Keane et al. 2015) and walrus (*Odobenus rosmarus*, Genbank: GCF_000321225.1) (Foote et al. 2015). Genome-wide average autosomal heterozygosity for the polar bear (*Ursus maritimus*, Genbank: GCF_000687225.1) (Liu et al. 2014), was taken from Westbury et al, 2018 (Westbury et al. 2018), while the following methods were implemented for the other species. To determine which scaffolds were most likely autosomal in origin, we found putative sex chromosome scaffolds for each of the species under investigation and removed them from future analyses. We found putative sex chromosome scaffolds in the narwhal, beluga, and bowhead whale by aligning the assembled genomes to the Cow X (Genbank: CM008168.2) and Human Y (Genbank: NC_000024.10) chromosomes. We found the putative sex chromosome scaffolds in the polar bear, and walrus by aligning the assembled genomes to the Human Y and the Dog X (Genbank: CM000039.3) chromosomes. Alignments were performed using satsuma synteny (Grabherr et al. 2010) and utilising default parameters.

We trimmed adapter sequences from the downloaded raw reads using skewer, mapped the trimmed reads to each respective reference genome using BWA v0.7.15 (Li & Durbin 2009) and the mem algorithm. We parsed the output and removed duplicates with samtools v1.6 (Li et al. 2009). Furthermore, to ensure comparability with previous heterozygosity estimates and to remove biases in heterozygosity levels that could arise due to different global coverages between the genomes of the individuals being investigated, we subsampled all of the resultant alignments down to 20x using samtools. We estimated the autosomal heterozygosity using sample allele frequencies in ANGSDv0.921 (Korneliussen et al. 2014), taking genotype likelihoods into account and specifying the following filters -minq 25 -minmapq 25 -uniqueOnly 1 -baq 1 -remove_bads 1 as was previously done in Westbury et al 2018 (Westbury et al. 2018). We computed the heterozygosity using ANGSD as it can overcome biases that may arise due to differential coverage across the genome. Instead of relying on direct SNP/genotype calling from the data, ANGSD uses genotype likelihoods data in downstream analyses and allows for the incorporation of statistical uncertainties into the analysis. This feature should reduce the biases caused by differential coverage across the genome.

The resultant values were compared alongside previously reported values from 10 other mammalian species (Westbury et al. 2018). We investigated heterozygosity in 500kb non-overlapping windows across the genomes of the five marine mammal species, using bedtools (Quinlan 2014). When plotting the results, we only considered windows from within the autosomes, scaffolds over 500kb in length, and windows with more than 70% data. Each window was treated individually and the percentage of heterozygous within each window was calculated. To investigate whether the heterozygosity results of the narwhal were a result of the quality of the genome, we mapped the short reads of our narwhal to the published beluga genome and repeated the above steps.

Finally, we investigated the distribution of heterozygosity across the genome, considering only autosomes and scaffolds longer than 500kb. This was done by independently calculating heterozygosity in five different partitions; exons, genes (exons +

introns), 10kb windows 10kb away, 20kb away, and 50kb away from the nearest protein-coding gene. We calculated variance in these results by randomly sampling 10% of the windows in each partition 100 times and plotting box plots using R. Using these 100 random samplings we additionally performed eight unpaired two sample t-tests per species to investigate the significance of differences between the different partitions. The comparisons included exons vs. autosomes, genes vs. autosomes, 10kb away vs. autosomes, 20kb away vs. autosomes, 50kb away vs. autosomes, exons vs. genes, genes vs, 10kb away, and genes vs. 20kb away. Differences were deemed significant by a p-value < 0.05.

Demographic history

We ran demographic analyses on diploid genomes from single individual species representatives of the narwhal, beluga, bowhead whale, walrus, and polar bear using a Pairwise Sequentially Markovian Coalescent model (PSMC)(Li & Durbin 2011). We called diploid genome sequences using samtools and bcftools (Narasimhan et al. 2016) specifying a minimum quality score of 20 and minimum coverage of 10. We removed scaffolds found to align to sex chromosomes in the previous step and scaffolds shorter than 100kb. We ran PSMC specifying atomic intervals previously shown to be suitable for human datasets (4+25*2+4+6) and performed 100 bootstrap replicates to investigate support for the resultant demography.

To estimate the mutation rate per generation for each species, we computed pairwise distances between closely related species, using a consensus base call in ANGSD and applying the filters -minQ 25 -minmapq 25 -uniqueonly 1 -remove_bads 1. Mutation rate per generation was calculated as follows: mutation rate = pairwise distance x generation time / 2 x divergence time. To estimate the narwhal and beluga mutation rates, short reads of both species were mapped to the narwhal genome, and mutation rate was calculated from the pairwise distances, assuming a divergence date of 5.5 Ma (Steeman et al. 2009). We assumed a narwhal generation time of 30 years and a beluga generation time of 32 years (Garde et al. 2015). To estimate the bowhead whale mutation rate, we downloaded short reads from the right whale (Genbank: SRR5665640) (Árnason et al. 2018) and mapped them to the bowhead whale genome. We calculated the mutation rate assuming a divergence date between the right whale and bowhead whale of 4.38 Ma (Árnason et al. 2018). We assumed a bowhead generation time of 35 years (Rooney et al. 2001). To estimate the walrus mutation rate, we mapped the northern fur seal (Genbank: SRR7278673) to the walrus genome and calculated the mutation rate assuming a divergence date between the walrus and the northern fur seal of 18 Ma (Higdon et al. 2007). We assumed a walrus generation time of 15 years (Andersen et al. 2009). For the polar bear, we used the previously published generational mutation rate of 1.825728e-08 and generation time of 11.2 years (Liu et al. 2014). Results and calculations can be seen in Supplemental tables S9 and S10.

Dating the end of gene flow between narwhal and beluga

To calculate when gene flow ceased between the narwhal and beluga, we used hPSMC (Cahill et al. 2016). To overcome any biases that may occur due to differences in reference qualities, we replicated this analysis twice, once with both species mapped to the narwhal genome and once with both species mapped to the beluga. We constructed haploid consensus sequences using ANGSD by considering the base with the highest effective depth, the following quality filters; -minQ 25, -minmapq 25, -uniqueonly 1, -remove_bads 1, -setMinDepthInd 10, and only considering autosomes and scaffolds over 100kb. These haploid consensus sequences were merged together using the hPSMC toolsuite into a pseudo diploid sequence, run through PSMC and plotted using a narwhal/beluga intermediate

mutation rate per generation of $1.6e^{-08}$ and an intermediate generation time of 31 years. From this output we estimated the pre-divergence Ne of the narwhal and beluga to be ~29,000 individuals. We ran simulations using this pre-divergence Ne with various divergence times between 1Ma and 2Ma in 50,000 year intervals using ms (Hudson 2002). Results were plotted and the simulations with an exponential increase in Ne closest to the real data, within 1.5x and 10x of the pre-divergence Ne, were taken as the time interval in which gene flow stopped.

References:

Andersen, L.W. et al., 2009. Genetic signals of historic and recent migration between sub-populations of Atlantic walrus *Odobenus rosmarus rosmarus* west and east of Greenland. *Endangered species research*, 9, pp.197–211.

Árnason, Ú. et al., 2018. Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Science advances*, 4(4), p.eaap9873.

Bushnell, B., 2014. BBTools software package. *URL http://sourceforge. net/projects/bbmap*.

Cahill, J.A. et al., 2016. Inferring species divergence times using pairwise sequential Markovian coalescent modelling and low-coverage genomic data. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 371(1699), p.20150138.

Foote, A.D. et al., 2015. Convergent evolution of the genomes of marine mammals. *Nature genetics*, 47(3), pp.272–275.

Garde, E. et al., 2015. Life history parameters of narwhals (*Monodon monoceros*) from Greenland. *Journal of mammalogy*, 96(4), pp.866–879.

Grabherr, M.G. et al., 2010. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* , 26(9), pp.1145–1151.

Grau, J.H. et al., 2018. Improving draft genome contiguity with reference-derived in silico mate-pair libraries. *GigaScience*, 7(5), p.giy029.

Gurevich, A. et al., 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* , 29(8), pp.1072–1075.

Higdon, J.W. et al., 2007. Phylogeny and divergence of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. *BMC evolutionary biology*, 7, p.216.

Holt, C. & Yandell, M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics*, 12, p.491.

Hudson, R.R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* . Available at: https://academic.oup.com/bioinformatics/article-abstract/18/2/337/225783.

Jiang, H. et al., 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC bioinformatics*, 15, p.182.

Jones, S.J.M. et al., 2017. The Genome of the Beluga Whale (*Delphinapterus leucas*). *Genes*, 8(12), p.378.

Keane, M. et al., 2015. Insights into the evolution of longevity from the bowhead whale genome. *Cell reports*, 10(1), pp.112–122.

Korf, I., 2004. Gene finding in novel genomes. *BMC bioinformatics*, 5, p.59.

Korneliussen, T.S., Albrechtsen, A. & Nielsen, R., 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC bioinformatics*, 15, p.356.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* , 25(16), pp.2078–2079.

Li, H. & Durbin, R., 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* , 25(14), pp.1754–1760.

Li, H. & Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357), pp.493–496.

Liu, S. et al., 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell*, 157(4), pp.785–794.

Luo, R. et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1), p.18.

NAMMCO, 2018. Report of the NAMMCO Global Review of Monodontids, Hillerød, Denmark. Available at: https://nammco.no/topics/sc-working-group-reports/.

Narasimhan, V. et al., 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* , 32(11), pp.1749–1751.

Parra, G. et al., 2009. Assessing the gene space in draft genomes. *Nucleic acids research*, 37(1), pp.289–297.

Parra, G., Bradnam, K. & Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* , 23(9), pp.1061–1067.

Paulino, D. et al., 2015. Sealer: a scalable gap-closing application for finishing draft genomes. *BMC bioinformatics*, 16, p.230.

Quinlan, A.R., 2014. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current protocols in bioinformatics*, 47, pp.11.12.1–34.

Rooney, A.P., Honeycutt, R.L. & Derr, J.N., 2001. Historical population size change of bowhead whales inferred from DNA sequence polymorphism data. *Evolution; international journal of organic evolution*, 55(8), pp.1678–1685.

Schmieder, R. & Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* , 27(6), pp.863–864.

Smit, A.F.A. & Hubley, R., 2008-2015. RepeatModeler Open-1.0. Available at: http://www.repeatmasker.org.

Smit, A.F.A., Hubley, R. & Green, P., 2013-2015. RepeatMasker Open-4.0. Available at: http://www.repeatmasker.org.

Stanke, M. & Waack, S., 2003. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* , 19, pp.215–225.

Steeman, M.E. et al., 2009. Radiation of extant cetaceans driven by restructuring of the oceans. *Systematic biology*, 58(6), pp.573–585.

Waterhouse, R.M. et al., 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Molecular biology and evolution*, 35(3), pp.543–548.

Westbury, M.V. et al., 2018. Extended and Continuous Decline in Effective Population Size Results in Low Genomic Diversity in the World's Rarest Hyena Species, the Brown Hyena. *Molecular biology and evolution*, 35(5), pp.1225–1237.