The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution

Fan, Guangyi; Zhang, Yaolei; Liu, Xiaochuan; Wang, Jiahao; Sun, Zeguo; Sun, Shuai; Zhang, He; Chen, Jianwei; Lv, Meiqi; Han, Kai

Total number of authors:
46

Published in:
Molecular Ecology Resources

Link to article, DOI:
10.1111/1755-0998.13003

Publication date:
2019

Document Version
Peer reviewed version

Article Type: From the Cover

The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution

Guangyi Fan[1,2,4,*], Yaolei Zhang[1,*], Xiaochuan Liu[1,*], Jiahao Wang[1,*], Zeguo Sun[1,3,*], Shuai Sun[1,*], He Zhang[1,*], Jianwei Chen[1,*], Meiqi Lv[1], Kai Han[1], Xiaoxuan Tan[1], Jie Hu[1], Rui Guan[1], Yuanyuan Fu[1], Shanshan Liu[1], Xi Chen[5], Qiwu Xu[1], Yating Qin[1], Longqi Liu[2], Jie Bai[2], Ou Wang[2], Jingbo Tang[2], Haorong Lu[3], Zhouchun Shang[2], Bo Wang[3], Guohai Hu[3], Xia Zhao[3], Yan Zou[2], Ao Chen[2], Meihua Gong[2], Wenwei Zhang[2], Simon Ming-Yuen Lee[4], Songhai Li[6], Junnian Liu[1], Zhen Li[7], Yishan Lu[8], Jamal S. M. Sabir[9], Mumdooh J. Sabir[9], Muhummadh Khan[9], Nahid H. Hajrah[9], Ye Yin[2,10], Karsten Kristiansen[10], Huanming Yang[2,11], Jian Wang[2,11], Xun Xu[1,2,3,#] & Xin Liu[1,2,3,#]

[1] BGI-Qingdao, BGI-Shenzhen, Qingdao, Shandong Province, 266555, China.
[2] BGI-Shenzhen, Shenzhen, 518083, China.
[3] China National GeneBank, BGI-Shenzhen, Shenzhen, 518120, China.
[4]State Key Laboratory of Quality Research of Chinese Medicine and Institute of Chinese Medical Sciences, University of Macau, Dama Road, Macau, China.
[5]Guangdong Pearl River Estuary Chinese White Dolphin National Nature Reserve Administration, Zhu Hai, Guangdong, China.
[6]Marine Mammal and Marine Bioacoustics Laboratory, Institute of Deep-sea Science and Engineering, Chinese Academy of Science, Sanya, China.
[7]Health and Family Planning Integrated Supervision Enforcement Bureau of Shinan District, Qingdao City, China.
[8]Guangdong Ocean University, Shenzhen, 518120, China.
[9]Department of Biological Sciences, King Abdulaziz University (KAU), Jeddah 21589, Saudi Arabia.

[10]Department of Biology, University of Copenhagen, Copenhagen, Denmark.
[11]James D. Watson Institute of Genome Sciences, Hangzhou 310058, China.

*These authors made equivalent contributions to this research.

[#]Corresponding authors (Xun Xu: xuxun@genomics.cn and Xin Liu: liuxin@genomics.cn)

Abstract

Marine mammals are important models for studying convergent evolution and aquatic adaption, thus reference genomes of marine mammals can provide evolutionary insights. Here, we present the first chromosome-level marine mammal genome assembly based on the data generated by the BGISEQ-500 platform, for a stranded female sperm whale (*Physeter macrocephalus*). Using this reference genome, we performed chromosome evolution analysis of sperm whale including constructing ancestral chromosomes, identifying chromosome rearrangement events and comparison with cattle chromosomes, which provides a resource for exploring marine mammal adaptation and speciation. We detected a high proportion of long interspersed nuclear elements (LINEs) and expanded gene families, and contraction of MHC region genes which were specific to sperm whale. By comparing to sheep and cattle, we conducted analysis of positively selected genes to identify gene pathways that may be related to adaptation to the marine environment. Further, we identified possible convergent evolution in aquatic mammals by testing for positively selected genes across three orders of marine mammals. In addition, we used publically available resequencing data to confirm a rapid decline in global population size in the Pliocene to Pleistocene transition. This study sheds light on the chromosome evolution and genetic mechanisms underpinning sperm whale adaptations, providing valuable resources for future comparative genomics.

Introduction

Sperm whale (*Physeter macrocephalus*) is the largest toothed whale (measuring up to 20.5 meters long and 57,000 kilograms in weight) (Perrin, Würsig, & Thewissen, 2009) with the largest brain among all animals (Marino, 2004). In addition, it is also one of the deepest and long-diving mammals (at depths of up to 3,000 meters and dive times up to 138 mins) (Schreer & Kovacs, 1997; Watwood, Miller, Johnson, Madsen, & Tyack, 2006). Sperm whales can adapt to hyperbaric and hypoxic environments by using their flexible ribcage which allows lung collapse, reducing nitrogen intake (Tyack, Johnson, Soto, Sturlese, & Madsen, 2006) (Kooyman & Ponganis, 1998). Sperm whales are distributed globally (Jaquet & Whitehead, 1996) and migrate seasonally to feeding and breeding grounds (Smith, Reeves, Josephson, & Lund, 2012). Because of its unique adaptations, recently, a draft genome has been published for a sperm whale stranded in the Gulf of Mexico (Warren et al., 2017). Draft genomes of bowhead whale (*Balaena mysticetus*), minke whale (*Balaenoptera acutorostrata*), antarctic minke whale (*Balaenoptera bonaerensis*), grey whale (*Eschrichtius robustus*), baiji (*Lipotes vexillifer*), white whale (*Delphinapterus leucas*), killer whale (*Orcinus orca*), bottlenose dolphin (*Tursiops truncates*), Yangtze finless porpoise (*Neophocaena asiaeorientalis*), harbour porpoise (*Phocoena phocoena*) are also available (Fan et al., 2018). However, to our knowledge, none of these currently available cetacean genomes have been assembled into chromosomes, which limits comparative genomic studies to some extent. Chromosome evolution such as whole genome duplication (WGD) and chromosome rearrangement accompanied with gene gain/loss and change of gene location is known to play important roles in evolving lineage-specific gene families, shaping organism's unique traits and even in speciation (Eichler & Sankoff, 2003; Kirkpatrick, 2010; Rockman, Skrovanek, & Kruglyak, 2010). For sperm whale, previous research has reported 21 pairs of chromosomes using fluorescence in situ hybridization (FISH), compared to other cetacean species with 22 pairs of chromosomes (Arnason, 1974). Chromosomal assembly will therefore provide a good foundation for studying the evolution of cetacean genomes.

In this study, we sampled a female sperm whale stranded near the bay area of Huizhou city in southern China on 14th March 2017. We extracted DNA from a muscle sample and conducted whole genome shotgun (WGS) sequencing, 10X Genomics Chromium™ sequencing as well as formaldehyde crosslinked Hi-C sequencing. In addition, we sequenced the transcriptome from RNA extracted from blood and muscle. We were therefore able to assemble and annotate a chromosome-level reference genome for this sperm whale, representing the first chromosome-level marine mammal genome. Investigating this reference genome, we were able to identify genome features including chromosome evolution and reshuffling, repeat content changes, regions under selection, and gene family expansion, which may relate to its adaptation.

Materials and Methods

DNA, RNA extraction, and library construction

See details in ***Supplementary materials****.*

Libraries sequencing and data filtering

A total of five WGS libraries were constructed with average insert sizes of 200 bp, 400 bp, 2 kb, 5 kb, and 10 kb as described in the supplementary information. The BGISEQ-500 sequencer was used to sequence these libraries and generate sequencing data from the paired-end libraries with read length of 100 bp and mate-pair libraries with read length of 50 bp, respectively. For the two transcriptome libraries with average insert size of ~250bp generated from blood and muscle samples**,** BGISEQ-500 was also used for yielding paired-end reads of 100 bp in length. For the 10X genomics library, paired-end reads of 150 bp in length were generated using BGISEQ-500 sequencer. The above raw reads with ratio of N (ambiguous base) higher than 5% and ratio of low-quality base (quality score less than 10) higher than 20% were removed by using SOAPnuke (v1.5.6) (Chen et al., 2017) with parameters 'filter -l 10 -q 0.2 -n 0.05 -Q 2 --misMatch 1 --matchRatio 0.4'. Duplicated reads which are identical in both ends were also removed by using SOAPnuke (v1.5.6) (Chen et al., 2017) with parameter '-d' to get final clean data. For the Hi-C library, BGISEQ-500 sequencer was again used for

yielding paired-end reads of 50 bp in length and raw reads were processed using the HiC-Pro pipeline (*see below*).

Genome assembly

Using sequencing data from WGS libraries (200 bp, 400 bp, 2 kb, 5 kb and 10 kb), we first assembled the genome using SOAP*denovo* (v2.04) (Luo et al., 2012) with parameters setting as '-K 49 –d 4 –D 4 –R', and gaps in scaffolds were then filled in using KGF (v1.19) and GapCloser (v1.10) (Luo et al., 2012) with default parameters. For assembly of the 10X Genomics Chromium library data, the clean fastq files were converted so as to be recognized by 10X Genomics Supernova (v1.2.0) (Weisenfeld, Kumar, Shah, Church, & Jaffe, 2017) using an in-house script. Reads were then *de novo* assembled using Supernova (v1.2.0) (Weisenfeld et al., 2017) with default parameters. At the 'mkoutput' stage in Supernova for outputting the assembly, we used the 'pseudohap' style and specified a minimum fasta record size of 100 bases. For the Hi-C data, the raw fastq files were processed using HiC-Pro (v2.8.0_devel) (Servant et al., 2015) pipeline with default parameters to get the valid reads. Then the WGS assembly result and 10X assembly result were then each anchored to chromosomes with the 3D-DNA pipeline (v170123) (Dudchenko et al., 2017) with parameters '-m haploid -s 4' using the Hi-C valid data. Finally, we employed GMcloser (v1.5.1) (Kosugi, Hirakawa, & Tabata, 2015) with parameters '-n 4 -b -mm 500 -mi 95 -ms 21 -c' to polish the WGS+Hi-C assembly result by using the 10X+Hi-C assembly result to get the final assembly.

Genome annotation of repeat elements and genes

Repeat elements were identified using both homology-based and *de novo* strategies. Firstly, RepeatMasker (v4.0.5) (Tarailo-Graovac & Chen, 2009) with parameters '-nolow -no_is -norna -engine ncbi' and RepeatProteinMasker (v4.0.5) with parameters '-engine ncbi -noLowSimple -pvalue 0.0001' were used to identify TEs at DNA and protein levels, respectively, by aligning against the Repbase (Bao, Kojima, & Kohany, 2015) database. Secondly, *de novo* repeat annotation was carried out using RepeatModeler (v1.0.8) with

default parameters and LTR-FINDER (v1.0.6) (Xu & Wang, 2007) (for long terminal repeats, LTRs) with default parameters. RepeatMasker was then used again with parameters '-nolow -no_is -norna -engine ncbi' to identify and classify repeat elements based on the *de novo* predicted repeats. Tandem Repeat Finder (v4.07) was used to find tandem repeats with parameters '-Match 2 -Mismatch 7 -Delta 7 -PM 80 -PI 10 -Minscore 50 -MaxPeriod 2000'. Finally, all the repeat elements identified above were further combined and classified using an in-house Perl script. The repeats were masked in the genome for further gene annotation. For annotation of repeat elements in other related species including the Gulf of Mexico (GM) sperm whale assembly (Warren et al., 2017), as well as bottlenose dolphin (*Tursiops truncatus*), minke whale (*Balaenoptera acutorostrata*), sheep (*Ovis aries*), wild boar (*Sus scrofa*) and cattle (*Bos taurus*), we downloaded genome sequences of these species from NCBI Reference Sequence Database (Release 86) and employed the same pipeline to get the repeat information. Then we summarized results generated by RepeatMasker and RepeatProteinMasker to get the sequence divergence of LINEs (defined by the two software) and get total lengths of different L1 elements (defined by the two software) in six mammal species. For constructing phylogenetic tree of LINE L1-1_Ttr, we extracted L1-EN domain which is conserved in LINEs (Repanas et al., 2007; Weichenrieder, Repanas, & Perrakis, 2004) through aligning LINE sequences to human L1-EN domain (https://www.ncbi.nlm.nih.gov/Structure/cdd/cddsrv.cgi?uid=197310) by using tBLASTn (v 2.2.26) (Altschul, Gish, Miller, Myers, & Lipman, 1990) with parameters '-m 9 -e 1e-5' followed by filtering with criteria 'identity > 60%, coverage> 95%' to do this analysis. Then MAFFT (v7.245) (Katoh & Standley, 2013) was used for multiple sequences alignment and FastTree (v2.1.10) (Price, Dehal, & Arkin, 2010) was used for constructing the tree with default parameters.

For gene annotation, we downloaded the protein sequences of 12 species from NCBI Annotation Release 105 (Pruitt, Tatusova, Brown, & Maglott, 2011) including baiji (*Lipotes vexillifer*), seal (*Leptonychotes weddellii*), killer whale (*Orcinus orca*), minke whale (*Balaenoptera acutorostrata*), manatee (*Trichechus manatus latirostris*), walrus (*Odobenus*

rosmarus divergens), bottlenose dolphin (*Tursiops truncatus*), wild boar (*Sus scrofa*), sheep (*Ovis aries*), cattle (*Bos taurus*), dog (*Canis lupus familiaris*) and human (*Homo sapiens*) for predicting homologous genes. Gene loci were identified by aligning these protein sequences to the sperm whale genome using tBLASTn (v 2.2.26) (Altschul et al., 1990) with an E-value cutoff of 1×10^{-5}. Gene models were predicted using GeneWise (v2.4.1) (Birney, Clamp, & Durbin, 2004) within these aligned gene loci with default parameters. For *de novo* annotation, AUGUSTUS (v3.1) (Stanke et al., 2006) and GENSCAN (v2009) (Burge & Karlin, 1997) were used with default parameters and the human data set was used as the training set of AUGUSTUS because human genes are much better annotated than any other close relative of sperm whale. For transcriptome genes annotation, we first assembled the transcriptome sequence using Trinity (v2.0.6) (Grabherr et al., 2011) with default parameters and the assembled sequences were mapped to the sperm whale genome using blat with default parameters. Finally, GLEAN (Elsik et al., 2007) with default parameters was used to integrate all the predicted gene models into a consensus gene set assessed by using BUSCO with vertebrata_odb9. The obtained gene set was mapped against Kyoto Encyclopedia of Genes and Genome (KEGG v84.0) (Kanehisa & Goto, 2000), Swissprot (v2017_09) (UniProt, 2012), TrEMBL (v2017_09) (Bairoch & Apweiler, 2000), and NR (v84) (Pruitt, Tatusova, & Maglott, 2006) databases using blastp with an E-value cutoff of 1×10^{-5} to find functionally similar genes. Gene motifs and domains were identified using InterProScan (v5.16-55.0) (Zdobnov & Apweiler, 2001) against ProDom (Bru et al., 2005), Pfam (Punta et al., 2012), SMART (Ponting, Schultz, Milpetz, & Bork, 1999), PANTHER (Mi et al., 2005), and PROSITE (Hulo et al., 2006). Gene ontology (GO) (Ashburner et al., 2000) terms were obtained from the InterPro entries.

Assessment of genome assembly

To reveal the improvement of the genome assembly, our assembly was aligned to the Gulf of Mexico (GM) assembly using Lastz (v1.02.00) (Harris, 2007) with parameters 'T=2 C=2 H=2000 Y=3400 L=6000 K=2200', and we calculated the alignment ratio of each scaffold of the GM assembly. We also detected insertions and deletions based on Lastz results. Then we

aligned our annotated genes to the genes in the GM assembly using BLASTP (v2.6.0+) (Altschul et al., 1990) with parameters 'e-value ≤ 1e-6, identity ≥80%' to find genes in the GM assembly which can be found in our consensus gene set. The remaining genes in our gene set were aligned to our transcriptome genes and homologous genes (see details in *Genome annotation of repeat elements and genes* section) to check whether they were supported by them.

Construction of ancestral chromosomes

Duplicate and syntenic regions in genomes serve as tracks of chromosome evolution, for example, whole genome duplication (WGD). And paralogous and orthologous gene pairs usually be used to identify duplicate and syntenic regions (Jaillon et al., 2004; Kellis, Birren, & Lander, 2004; Salse et al., 2008). To construct ancestral chromosomes of cattle and sperm whale, we firstly detected paralogous gene pairs in each species and orthologous gene pairs between these two species. We also detected paralogous gene pairs in human genome to support the conclusion of chromosome fusion events (*see details in the result section*). To detect paralogous and orthologous gene pairs, we firstly used blastp to perform protein alignment in each species and tblastx to do Coding DNA Sequence (CDS) alignment between species with E-value cutoff of 1e-5 to get the High Scoring Pairs (HSPs). Then we used software - Solar (v0.9.6) with default parameters to conjoin the HSPs. Next, we filtered the Solar result by both the query and target coverage and identity larger than 30%. Duplicate and syntenic regions were defined using MCSCAN (v0.8) (Tang et al., 2008) with parameter of "-a -e 1e-5 -s 5 -u 5" based on identified gene pairs. Based on the syntenic and duplicate relationships, ancestral chromosomes of sperm whale and cattle were reconstructed, and the rearrangement events were analyzed.

Synteny analysis with cattle

To visualize the concordance between the final sperm whale assembly and the cattle (*Bos taurus*) genome (Elsik, Tellam, & Worley, 2009), the 21 assembled sperm whale chromosomes were aligned to cattle chromosomes using Lastz (v1.02.00) (Harris, 2007) with

parameters the same as above. After filtering the aligned blocks shorter than 2 kb in length, we plotted the results using Circos (v0.69) (Krzywinski et al., 2009).

Gene family analysis

To define gene families, we firstly downloaded coding sequences of 13 species from NCBI Annotation Release 105 (Pruitt et al., 2011) and extracted the longest transcript for each gene. The 13 species include baiji (*Lipotes vexillifer*), seal (*Leptonychotes weddellii*), killer whale (*Orcinus orca*), minke whale (*Balaenoptera acutorostrata*), manatee (*Trichechus manatus latirostris*), walrus (*Odobenus rosmarus divergens*), dolphin (*Tursiops truncatus*), sheep (*Ovis aries*), cattle (*Bos taurus*), dog (*Canis lupus familiaris*), elephant (*Loxodonta africana*), opossum (*Monodelphis domestica*) and platypus (*Ornithorhynchus anatinus*). Gene families were identified using TreeFam (v4.0) (Li et al., 2006) with default parameters. Café (v2.1) (Hahn, Demuth, & Han, 2007) was used to define expansion and contraction of each gene family with default parameters. Single-copy gene orthologs were used to construct a phylogenetic tree. Each single-copy gene family was concatenated into a supergene for each species. Fourfold degenerate sites identified within each supergene were used to reconstruct the phylogenetic tree using PhyML (v3.0) (Grabherr et al., 2011), and the divergence time of species was estimated using MCMCtree with default parameters from the PAML package (Yang, 2007).

For KEGG and GO enrichment analysis, we first mapped the target genes to KEGG pathways and GO terms. Hypergeometric tests were performed to evaluate the significance of enriched genes and pathways using the whole genome as background.

For comparison of MHC region with cattle, we firstly used Lastz (v1.02.00) (Harris, 2007) to identify the MHC regions of sperm whale by aligning whole genome sequences of sperm whale to the cattle MHC regions (Takeshima & Aida, 2006). Then we extracted all genes in this region of sperm whale and cattle based on the gene function annotation information (the gene function annotation information of cattle comes from NCBI database). According to the gene function annotation result, we matched gene pairs and found the unique genes and lost

genes in two species. For verifying the lost genes in sperm whale, we employed Hisat2 (v2.0.4) (Pertea, Kim, Pertea, Leek, & Salzberg, 2016) with default parameters to do reads mapping to cattle genes using ~650M (~50X of the assembled genome) clean WGS sequencing read pairs (see details in *data filtering* section).

Positively selected genes detection

To identifying positively selected genes (PSGs) in the sperm whale genome, we selected sperm whale as the foreground branch with cattle and sheep as the background branch. Based on gene families of 13 species identified using TreeFam (see details in *gene family analysis* section), the CDS and protein sequences of single copy orthologous gene families for whale sperm, cattle and sheep were extracted. MUSCLE (v3.8.31) (Edgar, 2004) with parameter "-maxiters 2" was used for multiple sequence alignments of each single copy orthologous gene family of these three species, and Gblocks (v0.91b) (Castresana, 2000) was used to remove poorly aligned positions with parameters "-a=y -c=y w=y -t=c -e=gb1 -b4=5 -d=y". The PSGs were identified by comparing the alternative model (fix_omega = 1, omega = 1) to the null model (fix_omega = 0, omega = 1.5), and then the likelihood ratio tests (LRTs) were performed with a critical value of 3.84 at a 5% significance level using codeml in the PAML package (Yang, 2007). We used the false discovery rate (FDR) correction for multiple comparisons. For KEGG enrichment analysis, we first mapped the target genes to KEGG pathways, and then hypergeometric tests were performed to evaluate the significance of enriched pathways using the genes of sperm whale as background. The same analysis was performed to Cetacean (foreground: sperm whale, killer whale; background: cattle), Pinnipedian (foreground: walrus, seal, background: dog) and Sirenian (foreground: manatee, background: elephant) orders, respectively.

Population analysis

The raw data of five previously sequenced WGS (whole genome sequencing) individuals were downloaded from NCBI SRA database (SEY-1-Indian: SRX2447268, SRX2447275; SEY-2-Indian SRX2447270, SRX2447273; SC-1-Pacific: SRX2447271, SRX2447274; SC-

2-Pacific: SRX2447269, SRX2447272; GM-Atlantic: SRX220366, SRX220367) (Warren et al., 2017), and then converted to fastq format using "fastq-dump" in the sratoolkit package with default parameters. Adding the individual sequenced in our study, a total of 6 sperm whale individuals were used for population study. We removed reads containing greater than 50% low-quality bases (Q value ≤5), containing more than 5% unidentified (N) bases or containing sequencing adapter using SOAPnuke (v1.5.6) (Chen et al., 2017). The clean reads were mapped to our assembled sperm whale genome using BWA mem (v0.7.12-r1039) (Li & Durbin, 2010) with default parameters except the parameter "-M". The generated SAM files were converted into BAM files using SAMtools (v0.1.19-44428cd) (Guindon, Delsuc, Dufayard, & Gascuel, 2009), and then sorted by reference position using "SortSam.jar" in the picard package (v1.54). To improve the quality of alignment, local realignment was conducted using RealignerTargetCreator and IndelRealigner in GATK (v3.6) (Van der Auwera et al., 2013) with default parameters. SNPs were called using HaplotypeCaller and filtered using VariantFiltration with parameter "--genotypeFilterExpression 'DP < 3.0' --genotypeFilterName lt_3 --setFilteredGtToNocall --filterExpression "QD < 2.0 || FS > 60.0 || MQ <40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0"". SNPs were extracted using bcftools (v1.2, http://github.com/samtools/bcftools) for subsequent analysis.

The relatedness between 6 individuals was measured by calculating 1 minus the identity-by-state (IBS) matrix (i.e., a distance matrix) using PLINK (v1.90b) (Chang et al., 2015) with parameters "--distance 1-ibs flat-missing", and then the phylogenetic tree was constructed with neighbor joining method using "neighbor" in PHYLIP package (v3.69, http://evolution.genetics.washington.edu/phylip.html) with default parameters. To determine the individuals that were most closely related, principal component analysis was conducted on the top 10 principal components from the IBS matrix, calculated using PLINK (v1.90b) with parameter "--pca 10".

The history of change in effective population size was reconstructed using PSMC (v0.6.5-r67) (Li & Durbin, 2011). Firstly, diploid genome references for 6 individuals were constructed using samtools and bcftools call with "samtools mpileup -C30" and "vcfutils.pl vcf2fq -d 10 -D 100". Secondly, the demographic history was inferred using PSMC with parameters '-N25 -t15 -r5 -p 4+25*2+4+6' chosen following (Warren et al., 2017). The estimated generation time (g) and mutation rate per generation per site (μ) were set to 32 and 2.0e-8.

Results and Conclusion

The chromosome-level genome assembly of sperm whale

In order to obtain a high-quality reference genome assembly, we carried out different sequencing strategies on the new sequencing platform, BGISEQ-500 (Mak et al., 2017). In total, we obtained ~755 Gb sequencing data with sequence length ranging from 50 bp to 150 bp (**Supplementary Table 1** and **Supplementary Figure 1**). Combining the WGS data, the 10X data, as well as the Hi-C data (**Figure 1A**), we obtained a genome assembly with the total length of 2.58Gb (close to the estimated genome size, **Supplementary Figure 2**), contig N50 of 48.81 kb and scaffold N50 of 121.90 Mb, and 94.34% of the assembled genome was anchored onto 21 chromosomes (**Figure 1B** and **Supplementary Table 2**). We found high consistency of the length (correlation coefficient: ~0.93) between the chromosomes and the karyotypes (Arnason, 1974) (**Figure 1C**), indicating good quality of the chromosome anchoring. With this genome assembly, we annotated 20,740 protein-coding genes (**Supplementary Table 3**), which was close to the average gene number of the 17 published aquatic mammals (22,518) (**Supplementary Table 4**). For the 20,740 genes, 90.19% of them can be functionally annotated (**Supplementary Table 5**) and 94.1% of the 2,586 BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) gene models can be found in the annotated genes (**Supplementary Table 6**), indicating the high accuracy of our gene prediction.

We then compared the genome assembly to the published genome assembly of a sperm whale stranded in the Gulf of Mexico (GM assembly) sequenced using an Illumina sequencing platform (Warren et al., 2017). We found high coverage (~99.29%) of the GM assembly (~2.28Gb) in our genome assembly (**Figure 1D** and **Supplementary Figure 3**), with 341,881 insertions and deletions (ranging from 1 bp to 143 bp, 1.18 Mb in total, **Supplementary Figure 4** and **5**), and 981 inversions (ranging from 185 bp to 76 kb). Further looking into the comparison of repetitive sequences between the GM assembly (~921.64 Mb) and our genome assembly (~938.82 Mb), we found similar repetitive sequence distribution in the aligned regions, while in the unaligned regions, we found more repetitive sequences in our genome assembly (**Supplementary Figure 6**) indicating better coverage of repeat regions of our genome assembly. For the protein-coding genes, we found that most of the genes (17,672 out of 18,563) in the GM assembly can be found in our gene set and most of the remaining genes in our gene set (2,534 out of 3,140) were supported by the transcriptome data or homologous genes, again indicating the high quality of our gene annotation.

Specific LINEs burst in sperm whale

Despite similar repeat content, we found substantially higher proportions of long interspersed nuclear elements (LINEs) (85.38% in all repeats) in sperm whale comparing to the other species (less than 65%) (**Supplementary Table 7**). We further calculated the divergence rate of LINE sequences in three marine species (sperm whale, minke whale and bottlenose dolphin) and three terrestrial species (cattle, sheep and wild boar). We found similar LINE divergence patterns with two peaks in these six species, with the first peaks at the divergence rates of 5~8% (**Supplementary Figure 7**). Nevertheless, the divergence rates indicated by the second peaks were found to be relatively lower in marine mammals compared to terrestrial mammals. This reflects a slower rate of evolution compared to terrestrial mammals which is consistent with the previous report of a slower molecular clock in whale (Jackson et al., 2009).

Furthermore, we described the abundance of L1-2_Ttr and L1-1_Ttr (subtypes of LINE) present in the genomes of marine mammals compared to terrestrial mammals (**Supplementary Figure 8**). From the phylogenetic analysis, we found L1-1_Ttr of sperm whale were abundant in different types of L1-1_Ttr, while that of minke whale were just abundant in one type (**Supplementary Figure 9**). We also identified several types of L1-1_Ttr to be only found in marine species. L1 can lead to homologous recombination that can result in epigenetic dysregulation and cause some genetic defects and diseases (Burwinkel & Kilimann, 1998; Segal et al., 1999). Thus, abundance of L1 in sperm whales and other aquatic species might have been important for their evolution.

Chromosomal evolution of sperm whale

Comparing to cattle with 30 pairs of chromosomes and other cetacean species with 22 pairs of chromosomes , Physeteridae have 21 pairs of chromosomes (Arnason, 1974). Our chromosome-level genome assembly makes it possible to analyze chromosome evolution from terrestrial mammals to marine mammals. Previous study (Nakatani, Takeda, Kohara, & Morishita, 2007) demonstrated that the chromosome number of the ancestor of vertebrates was 23 after the first whole genome duplication (WGD) event and the karyotype of the ancestor of eutherians was also 23 after the second WGD event (Ferguson-Smith & Trifonov, 2007). To investigate the chromosome evolution of sperm whale, we constructed the ancestral chromosomes of cattle and sperm whale using human as an outgroup. First, we identified 3,997, 4,285 and 3,033 conserved paralogous genes and then detected 89, 72 and 59 large paralogous blocks including 1,085, 1,090 and 581 gene pairs in human, cattle and sperm whale, respectively. We found that most of these conserved paralogous gene pairs (70.51%, 86.15% and 69.54% of all the conserved paralogous gene pairs in the three species, respectively) (**Supplementary Tables 8**-**10**) had both genes located on the same chromosome, indicating that after the second WGD, the duplicated chromosome pair were fused, instead of fusing other chromosomes. In addition, we also identified 14 major inter-chromosomal rearrangements during the ancestral chromosome evolutionary process (**Figure 2** and **Supplementary Table 11**, green).

Then, we identified orthologous genes among human, cattle and sperm whale to infer shared duplications which may be conserved from the common ancestor. In order to establish ancestral chromosomes for cattle and sperm whale, with both the orthologous and paralogous gene information, we obtained 58 shared duplications including 568 paralogous gene pairs between cattle and sperm whale (**Supplementary Table 11**), which should result from 23 ancestral chromosomes, the same as FISH-based estimation of eutherian chromosomes (Ferguson-Smith & Trifonov, 2007; Wienberg, 2004) (**Supplementary Table 11**, yellow). In this way, we found five chromosome fissions happened in cattle, while four fusion events happened in sperm whale which resulted in the chromosome number differences between them (**Figure 2**). Our analyses illustrated the chromosome evolution process for mammals, and especially for sperm whale, paving a way for better understanding of marine mammal evolution.

To further explore the chromosome evolution of cattle and sperm whale, we aligned the sperm whale genome against the cattle genome to find a mean coverage of 97.15% (**Figure 3A** and **Supplementary Table 12**). Eleven chromosomes of sperm whale could be uniquely mapped to single chromosomes of cattle, while the other ten chromosomes could be mapped to two or more chromosomes (**Figure 3B** and **Supplementary Figure 10** and **11**). In this way, we identified 30 chromosomal reshuffling events in sperm whale, which can be confirmed by sequencing data (**Supplementary Figure 12** and **13**). Within these reshuffling events, we identified four protein-coding genes and four pseudogenes located in the 2 kb flanking region of breakpoints (**Supplementary Figure 14** and **Supplementary Table 13**), such as *IMPA1*, lacking of the start codon in sperm whale, involved in salinity adaptation and that has been suggested to protect inositol in various tissues of the euryhaline eel (*Anguilla anguilla*) exposed to hypertonic environments (Kalujnaia, McVee, Kasciukovic, Stewart, & Cramb, 2010). We also identified the enhancers of the regulatory binding region of four protein-coding genes located on the breakpoint regions, inferring that the breakpoints probably contribute the regulation of regulatory elements. Thus, these reshuffling events

likely play a role in the adaptation of sperm whale by influencing the function of genes or regulatory elements.

Sperm whale specific gene family evolution

Expansion and contraction of gene families are thought to be important in adaptive phenotypic diversification (Hahn, De Bie, Stajich, Nguyen, & Cristianini, 2005). We identified 1,168 expanded and 2,211 contracted gene families (*p*-value <= 0.05) (**Supplementary Figure 15**) in sperm whale, among which, 502 and 396 gene families were specifically significantly expanded and contracted in sperm whale (*p*-value<0.01). Of those specifically expanded genes families, we noted five gene families probably with important functions (**Supplementary Figure 16**), which were involved in stabilization of newly synthesized DNA (K01581, *ODC*, Ornithine decarboxylase), bone development (K04673, *BMPR1A*, *bone morphogenetic protein receptor 1a*), prevention of the production of iron-catalyzed reactive oxygen species (K00522, *FTH*, ferritin heavy chain; *K13625*, *FLT*, ferritin light chain), modulation of responses to hypoxic, oxidative, and osmotic stresses (K09667, *OGT*, O-linked N-acetylglucosamine (GlcNAc) transferase), regulation of peroxide levels (K13279, *PRDX1*, *peroxiredoxin 1*). Of these expanded gene families, *OGT* and *PRDX1* had also been reported in minke whale (Yim et al., 2014). Those expanded genes families in sperm whale may be linked to the special ability to adapt to deep and long diving (Schreer & Kovacs, 1997).

For the contracted gene families, we particularly analyzed gene families in the major histocompatibility complex (MHC) region, which are thought to be important disease-related genes. We identified ~6.0 Mb MHC regions on chromosome 18 and chromosome 21 in sperm whale by mapping the ~5.4 Mb MHC I, II and III regions from cattle chromosome 23 (Takeshima & Aida, 2006) (**Supplementary Figure 17**). We compared the MHC genes and found 136 gene families shared between sperm whale and cattle, while 73 genes were missing and 27 genes were unique in sperm whale. Among the 136 shared gene families, we found five copies of *H2B* type 2-E (*H2B*) in sperm whale comparing to only one *H2B* gene in

cattle (**Figure 3C**). *H2B* plays a vital role in DNA replication and repair and is associated with the rate of mRNA elongation by RNA polymerase II, thereby increasing the rate of gene expression (Fuchs, Hollander, Voichek, Ast, & Oren, 2014). For 73 genes missing in sperm whale (**Supplementary Table 14**), forty of which were olfactory receptors (*Olfr*), including *Olfr12D*, *Olfr2G*, *Olfr2B*, *Olfr2W* and *Olfr10C* (**Supplementary Figure 18**), which are thought to be involved in olfaction-driven mate selection (Spehr et al., 2006; Younger et al., 2001). We also investigated functions of the 27 genes unique to sperm whale (**Supplementary Figure 19** and **Supplementary Table 15**). For example, SpermWhale_09225 was a homologous gene of Ubiquilin 1, which has been reported to mediate degradation of proteins involved in stress response and neurotransmission, reflecting the functional importance of these unique genes.

Positively selected genes

In addition to expansion and contraction of gene families, genes which have undergone positive selection commonly contribute to adaptive phenotypic evolution and adaptation. We identified positively selected genes (PSGs) in the sperm whale genome by comparing to cow and sheep. Among the 8,674 one-to-one orthologs, 1,306 genes were identified to be PSGs (FDR<0.1). Some of these PSGs have already been reported in marine mammals (**Supplementary Table 16**), such as *SLC16A1* (Solute Carrier Family 16 Member 1; related to adaptation to long dives) and *GPR97* (G212 Protein-Coupled Receptor 97) (Foote et al., 2015). We performed KEGG enrichment analysis of these PSGs and identified six KEGG pathways that were significantly enriched, which may be related to the adaption to the marine environment (**Supplementary Table 17**) (Aedo et al., 2015; Lai et al., 2016). In addition, we performed branch-site likelihood ratio tests for the three orders of marine mammals to identify possible convergent evolution in aquatic mammals. We compared sperm whale and killer whale (*Orcinus orca*) to cattle for the Cetacean order, walrus (*Odobenus rosmarus*) and seal (*Phoca vitulina*) to dog (*Canis lupus familiaris*) for the Pinnipedian order, and manatee (*Trichechus manatus latirostris*) to elephant (*Loxodonta africana*) for the Sirenian order and

identified 444, 676 and 1,107 PSGs (FDR < 0.01) respectively. 111 PSGs were shared in at least two branches, and significantly enriched in hematopoietic cell lineage and apoptosis (**Supplementary Table 18**).

Evolution of sperm whale population

In order to analyze the population evolution of sperm whale, we included the resequencing data of five sperm whales from a previous study (Warren et al., 2017). Mapping the sequencing data of the five sperm whales to our reference genome, we identified ~8.47 million SNPs in total, with a diversity level of 0.00136, comparing to 0.0268 in Indian cattle (Sharma et al., 2015), 0.0009 in killer whale (Foote et al., 2016) and 0.0008 in finless porpoises (Zhou et al., 2018). Using the identified variations, we first inferred the population demographic changes using PSMC, which showed similar population history as illustrated previously (Warren et al., 2017) (**Supplementary Figure 20A**) with rapid decline in population size during the Pliocene to Pleistocene transition and increases afterwards. Also, we found a similar pattern estimated using data of different individuals, including the one we sequenced here which was sampled in the west Pacific. Then, we further analyzed the relationships between the six individuals. From the phylogenetic tree (**Supplementary Figure 20B**), the individual we sequenced was more closely related to the two samples from the Indian Ocean (Seychelles), while were quite different from the two samples from the east Pacific Ocean (Gulf of California). Furthermore, we conducted PCA to reveal relationships of these individuals (**Supplementary Figure 20C**). For the first principle component, the individuals from Indian Ocean showed higher diversity, while the second principle component distinguished the individual from Atlantic Ocean (Gulf of Mexico) from the others. Thus, overall, we found close relationships between the west Pacific Ocean individual and the Indian Ocean individuals compared to the east Pacific Ocean individuals, which may also be caused by the seasonal migration phenomenon in sperm whales (Smith et al., 2012).

Conclusion

As the first large-genome species sequenced and assembled using BGISEQ-500 sequencing data, the assembled sperm whale genome was highly contiguous and of good quality. In this study, we explored several assembly strategies (WGS+Hi-C, 10X+Hi-C and WGS+Hi-C polished by using 10X+Hi-C) using second generation sequencing including hierarchical whole genome shotgun and 10X Genomics and Hi-C. We proved the efficiency of constructing high quality, chromosomal-level reference genomes under these strategies. Using the genome assembly of sperm whale, we were able to investigate genome evolution at chromosome, repeat content, gene family and gene scales. We identified genomic events putatively related to the adaptation of sperm whale to water environments, and also to phenotypic variation during evolution at chromosomal level for the first time. Given the importance of understanding chromosome evolution in order to interpret a lot of biological questions, such as identifying the sex-determining chromosome, this first chromosomal level genome will be an important resource for future genomic research of marine mammals. By using a Hi-C strategy, chromosomal level genomes can be obtained more easily than before. The availability of more and more genomes at chromosomal level will enable macro-evolutionary analysis, for instance, inferring the chromosome evolution of marine mammals and relationships with their terrestrial ancestors. Thus, the results of this study provide valuable resources for future genomic, ecological and evolutionary studies of aquatic or marine mammals, and will also serve as reference for further genomic studies using the BGISEQ-500 sequencing platform.

Supplementary Information is available in the online version of the paper.

Competing Interests: The authors declare that they have no competing interests.

Acknowledgements

This work was supported in part by the Shenzhen Municipal Government of China Peacock Plan NO. KQTD20150330171505310 and Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011). We would also like to thank Dr. Varshney Rajeev from International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) to help us revising the manuscript and provide valuable suggestions.

Authors Contributions

X.X. and X.L. designed and managed this project. S.L., Q.X., J.B., O.W., J.T., H.L., B.W., G.H., Y.Q., L.L, Z.S. and X.Z. performed sample preparation and sequencing. G.F., Z.S., X.L., and Y.F. performed genome assembly. G.F., Z.S., and R.G. performed genome annotation. G.F., Z.S., S.S., J.W., H.Z., Y.Z., X.L., M.L., J.H., J.L. and K.H. performed genetic analysis. J.C. and X.T. performed metagenomics analysis. G.F. wrote the paper with contributions from Z.S., S.S, J.W., H.Z., Y.Z., J.C., M.J.S, M.K, N.H.H and X.L. All authors helped with the interpretation of data.

Data Accessibility

The final sperm whale genome assembly has been deposited at NCBI under project PRJNA411766 and CNGB database with accession number CNA0002349. All the sequencing data are also available under this BioProject with accession number ERS2373129 and ERS2373131. All custom scripts are available on Github (https://github.com/fish-school/src-for-spermwhale-project).

Figure legends

Figure 1. Assembly of the sperm whale genome. A) Assembly strategy. We first conducted genome assembly using WGS data and 10X Genomics respectively, then used the Hi-C data to anchor these two scaffold level assemblies to chromosomes respectively, and finally combined the two chromosomal assemblies to get the final genome assembly. The histogram shows the statistics of total length, scaffold N50 and contig N50. **B**) 21 chromosome contact maps of sperm whale at 125 kb resolution. The blocks represent the contact between one location and the other locations. The color reflects the intensity of each contact, for which the deeper color represents the higher intensity. Chromosomes were marked by blue arrows. **C)** The correlation between the length of karyotype and chromosome assembled in sperm whale. The length of chromosomes and karyotype were arranged from small to large respectively and corresponded to each other, and the correlation coefficient was calculated as 0.93. **D)** The coverage rate and scaffolds length distribution of the Gulf of Mexico (GM) assembly compared to our assembly. The GM assembly scaffolds with coverage rate >= 70% are shown. The top histogram indicates the distribution of coverage rate, the right histogram indicates the distribution of scaffolds length, and the heat map indicates that the majority of GM assembly scaffolds were fragmented but with high coverage rate.

Figure 2. **Reconstruction of ancestral chromosomes of cattle and sperm whale**. The figure displays the distribution of ancestral chromosome segments in cattle and sperm whale genomes, including 14 inter-chromosome rearrangements, fission and fusion events in cattle and sperm whale. Capital letters A-W and corresponding color rectangles below represent 23 reconstructed ancestral chromosomes. These chromosomes next experienced a second whole genome duplication (WGD), which was then followed by self-fusion and rearrangements. 53-59 MYA refers to the divergence time of cattle and sperm whale. Black arrows: chromosome rearrangements of ancestral chromosomes of cattle and sperm whale. Color arrows in cattle box: chromosomes fission events. Green arrows in sperm whale box: chromosomes fusion events.

Figure 3. Genome features of sperm whale and comparison of MHC region with cattle. **A**) Genome comparison between sperm whale and cattle. Green rectangles on the left of the circle represent sperm whale chromosomes, and rectangles of other colors on the right represent cattle chromosomes. **B**) Three types of collinear relationship between sperm whale and cattle. The upper red horizontal lines represent cattle chromosomes, the lower horizontal lines in various colors represent sperm whale chromosomes, and the lines between the two horizontal lines link the alignment blocks. The blue points represent the breakpoint regions. Yellow rectangles represent the gene density in 1Mb windows, and blue rectangles represent the repeat elements density in the same windows. The deeper the color, the higher the density. **C**). Detailed comparison of MHC genes between sperm whale and cattle. The black, green, and red rectangles represent genes shared among cattle and sperm whale, cattle specific genes, and sperm whale specific genes, respectively. The arrows show orientation of the chromosome from centromere to telomere.

References

Aedo, J. E., Maldonado, J., Aballai, V., Estrada, J. M., Bastias-Molina, M., Meneses, C., . . . Valdés, J. A. (2015). mRNA-seq reveals skeletal muscle atrophy in response to handling stress in a marine teleost, the red cusk-eel (Genypterus chilensis). *BMC genomics, 16*(1), 1024.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410. doi:10.1016/S0022-2836(05)80360-2

Arnason, U. (1974). Comparative chromosome studies in Cetacea. *Hereditas, 77*(1), 1-36.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., . . . Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet, 25*(1), 25-29. doi:10.1038/75556

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res, 28*(1), 45-48.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA, 6*, 11. doi:10.1186/s13100-015-0041-9

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. *Genome Res, 14*(5), 988-995. doi:10.1101/gr.1865504

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., & Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res, 33*(Database issue), D212-215. doi:10.1093/nar/gki034

Burge, C., & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J Mol Biol, 268*(1), 78-94. doi:10.1006/jmbi.1997.0951

Burwinkel, B., & Kilimann, M. W. (1998). Unequal homologous recombination between LINE-1 elements as a mutational mechanism in human genetic disease. *J Mol Biol, 277*(3), 513-517. doi:10.1006/jmbi.1998.1641

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution, 17*(4), 540-552.

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience, 4*, 7. doi:10.1186/s13742-015-0047-8

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., . . . Li, Z. (2017). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *GigaScience, 7*(1), gix120.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., . . . Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science, 356*(6333), 92-95. doi:10.1126/science.aal3327

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research, 32*(5), 1792-1797.

Eichler, E. E., & Sankoff, D. (2003). Structural dynamics of eukaryotic chromosome evolution. *Science, 301*(5634), 793-797.

Elsik, C. G., Mackey, A. J., Reese, J. T., Milshina, N. V., Roos, D. S., & Weinstock, G. M. (2007). Creating a honey bee consensus gene set. *Genome biology, 8*(1), 1.

Elsik, C. G., Tellam, R. L., & Worley, K. C. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science, 324*(5926), 522-528.

Fan, G., Chen, J., Jin, T., Shi, C., Du, X., Zhang, H., . . . Liu , X. L. (2018). The Report of Marine Life Genomic Research. *Preprints*. doi:10.20944/preprints201812.0156.v1

Ferguson-Smith, M. A., & Trifonov, V. (2007). Mammalian karyotype evolution. *Nature Reviews Genetics, 8*(12), 950.

Foote, A. D., Liu, Y., Thomas, G. W., Vinař, T., Alföldi, J., Deng, J., . . . Joshi, V. (2015). Convergent evolution of the genomes of marine mammals. *Nature genetics, 47*(3), 272.

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., . . . Martin, M. D. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature communications, 7*, 11693.

Fuchs, G., Hollander, D., Voichek, Y., Ast, G., & Oren, M. (2014). Cotranscriptional histone H2B monoubiquitylation is tightly coupled with RNA polymerase II elongation rate. *Genome research, 24*(10), 1572-1583.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., . . . Regev,

A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol, 29*(7), 644-652. doi:10.1038/nbt.1883

Guindon, S., Delsuc, F., Dufayard, J.-F., & Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *Bioinformatics for DNA sequence analysis*, 113-137.

Hahn, M. W., De Bie, T., Stajich, J. E., Nguyen, C., & Cristianini, N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome research, 15*(8), 1153-1160.

Hahn, M. W., Demuth, J. P., & Han, S.-G. (2007). Accelerated rate of gene gain and loss in primates. *Genetics*.

Harris, R. S. (2007). *Improved pairwise alignment of genomic DNA*: The Pennsylvania State University.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., De Castro, E., Langendijk-Genevaux, P. S., . . . Sigrist, C. J. (2006). The PROSITE database. *Nucleic Acids Res, 34*(Database issue), D227-230. doi:10.1093/nar/gkj063

Jackson, J. A., Baker, C. S., Vant, M., Steel, D. J., Medrano-Gonzalez, L., & Palumbi, S. R. (2009). Big and slow: phylogenetic estimates of molecular evolution in baleen whales (suborder mysticeti). *Mol Biol Evol, 26*(11), 2427-2440. doi:10.1093/molbev/msp169

Jaillon, O., Aury, J.-M., Brunet, F., Petit, J.-L., Stange-Thomann, N., Mauceli, E., . . . Bernot, A. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature, 431*(7011), 946.

Jaquet, N., & Whitehead, H. (1996). Scale-dependent correlation of sperm whale distribution with environmental features and productivity in the South Pacific. *Marine ecology progress series*, 1-9.

Kalujnaia, S., McVee, J., Kasciukovic, T., Stewart, A. J., & Cramb, G. (2010). A role for inositol monophosphatase 1 (IMPA1) in salinity adaptation in the euryhaline eel (Anguilla anguilla). *FASEB J, 24*(10), 3981-3991. doi:10.1096/fj.10-161000

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution, 30*(4), 772-780.

Kellis, M., Birren, B. W., & Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature, 428*(6983), 617.

Kirkpatrick, M. (2010). How and why chromosome inversions evolve. *PLoS biology, 8*(9), e1000501.

Kooyman, G., & Ponganis, P. (1998). The physiological basis of diving to depth: birds and mammals. *Annual Review of Physiology, 60*(1), 19-32.

Kosugi, S., Hirakawa, H., & Tabata, S. (2015). GMcloser: closing gaps in assemblies accurately with a likelihood-based selection of contig or long-read alignments. *Bioinformatics, 31*(23), 3733-3741.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M.

A. (2009). Circos: an information aesthetic for comparative genomics. *Genome Res, 19*(9), 1639-1645. doi:10.1101/gr.092759.109

Lai, K. P., Li, J. W., Tse, A. C. K., Cheung, A., Wang, S., Chan, T. F., . . . Wu, R. S. S. (2016). Transcriptomic responses of marine medaka's ovary to hypoxia. *Aquatic Toxicology, 177*, 476-483.

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Heriche, J. K., Osmotherly, L., . . . Durbin, R. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res, 34*(Database issue), D572-580. doi:10.1093/nar/gkj118

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics, 26*(5), 589-595. doi:10.1093/bioinformatics/btp698

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

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., . . . Liu, Y. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience, 1*(1), 18.

Mak, S. S. T., Gopalakrishnan, S., Caroe, C., Geng, C., Liu, S., Sinding, M. S., . . . Gilbert, M. T. P. (2017). Comparative performance of the BGISEQ-500 vs Illumina HiSeq2500 sequencing platforms for palaeogenomic sequencing. *Gigascience, 6*(8), 1-13. doi:10.1093/gigascience/gix049

Marino, L. (2004). Cetacean brain evolution: multiplication generates complexity. *International Journal of Comparative Psychology, 17*(1).

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., . . . Thomas, P. D. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res, 33*(Database issue), D284-288. doi:10.1093/nar/gki078

Nakatani, Y., Takeda, H., Kohara, Y., & Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res, 17*(9), 1254-1265. doi:10.1101/gr.6316407

Perrin, W. F., Würsig, B., & Thewissen, J. (2009). *Encyclopedia of marine mammals*: Academic Press.

Pertea, M., Kim, D., Pertea, G. M., Leek, J. T., & Salzberg, S. L. (2016). Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc, 11*(9), 1650-1667. doi:10.1038/nprot.2016.095

Ponting, C. P., Schultz, J., Milpetz, F., & Bork, P. (1999). SMART: identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res, 27*(1), 229-232.

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one, 5*(3), e9490.

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2011). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic acids research, 40*(D1), D130-D135.

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2006). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.

Nucleic acids research, 35(suppl_1), D61-D65.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., . . . Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Res, 40*(Database issue), D290-301. doi:10.1093/nar/gkr1065

Repanas, K., Zingler, N., Layer, L. E., Schumann, G. G., Perrakis, A., & Weichenrieder, O. (2007). Determinants for DNA target structure selectivity of the human LINE-1 retrotransposon endonuclease. *Nucleic acids research, 35*(14), 4914-4926.

Rockman, M. V., Skrovanek, S. S., & Kruglyak, L. (2010). Selection at linked sites shapes heritable phenotypic variation in C. elegans. *Science, 330*(6002), 372-376.

Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U. M., . . . Feuillet, C. (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *The Plant Cell, 20*(1), 11-24.

Schreer, J. F., & Kovacs, K. M. (1997). Allometry of diving capacity in air-breathing vertebrates. *Canadian Journal of Zoology, 75*(3), 339-358.

Segal, Y., Peissel, B., Renieri, A., de Marchi, M., Ballabio, A., Pei, Y., & Zhou, J. (1999). LINE-1 elements at the sites of molecular rearrangements in Alport syndrome-diffuse leiomyomatosis. *Am J Hum Genet, 64*(1), 62-69. doi:10.1086/302213

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C. J., Vert, J. P., . . . Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol, 16*, 259. doi:10.1186/s13059-015-0831-x

Sharma, R., Kishore, A., Mukesh, M., Ahlawat, S., Maitra, A., Pandey, A. K., & Tantia, M. S. (2015). Genetic diversity and relationship of Indian cattle inferred from microsatellite and mitochondrial DNA markers. *BMC Genet, 16*, 73. doi:10.1186/s12863-015-0221-0

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics, 31*(19), 3210-3212.

Smith, T. D., Reeves, R. R., Josephson, E. A., & Lund, J. N. (2012). Spatial and seasonal distribution of American whaling and whales in the age of sail. *PloS one, 7*(4), e34905.

Spehr, M., Kelliher, K. R., Li, X.-H., Boehm, T., Leinders-Zufall, T., & Zufall, F. (2006). Essential role of the main olfactory system in social recognition of major histocompatibility complex peptide ligands. *Journal of Neuroscience, 26*(7), 1961-1970.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., & Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res, 34*(Web Server issue), W435-439. doi:10.1093/nar/gkl200

Takeshima, S. N., & Aida, Y. (2006). Structure, function and disease susceptibility of the bovine major histocompatibility complex. *Animal Science Journal, 77*(2), 138-150.

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., & Paterson, A. H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome research*, gr. 080978.080108.

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements

in genomic sequences. *Curr Protoc Bioinformatics, Chapter 4*, Unit 4 10. doi:10.1002/0471250953.bi0410s25

Tyack, P. L., Johnson, M., Soto, N. A., Sturlese, A., & Madsen, P. T. (2006). Extreme diving of beaked whales. *Journal of Experimental Biology, 209*(21), 4238-4253.

UniProt, C. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res, 40*(Database issue), D71-75. doi:10.1093/nar/gkr981

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., Del Angel, G., Levy-Moonshine, A., . . . DePristo, M. A. (2013). From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics, 43*, 11 10 11-33. doi:10.1002/0471250953.bi1110s43

Warren, W. C., Kuderna, L., Alexander, A., Catchen, J., Perez-Silva, J. G., Lopez-Otin, C., . . . Wise, J. P., Sr. (2017). The Novel Evolution of the Sperm Whale Genome. *Genome Biol Evol, 9*(12), 3260-3264. doi:10.1093/gbe/evx187

Watwood, S. L., Miller, P. J., Johnson, M., Madsen, P. T., & Tyack, P. L. (2006). Deep-diving foraging behaviour of sperm whales (Physeter macrocephalus). *J Anim Ecol, 75*(3), 814-825. doi:10.1111/j.1365-2656.2006.01101.x

Weichenrieder, O., Repanas, K., & Perrakis, A. (2004). Crystal structure of the targeting endonuclease of the human LINE-1 retrotransposon. *Structure, 12*(6), 975-986.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Res, 27*(5), 757-767. doi:10.1101/gr.214874.116

Wienberg, J. (2004). The evolution of eutherian chromosomes. *Curr Opin Genet Dev, 14*(6), 657-666. doi:10.1016/j.gde.2004.10.001

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res, 35*(Web Server issue), W265-268. doi:10.1093/nar/gkm286

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution, 24*(8), 1586-1591.

Yim, H.-S., Cho, Y. S., Guang, X., Kang, S. G., Jeong, J.-Y., Cha, S.-S., . . . Kwon, K. K. (2014). Minke whale genome and aquatic adaptation in cetaceans. *Nature genetics, 46*(1), 88.

Younger, R. M., Amadou, C., Bethel, G., Ehlers, A., Lindahl, K. F., Forbes, S., . . . Trowsdale, J. (2001). Characterization of clustered MHC-linked olfactory receptor genes in human and mouse. *Genome research, 11*(4), 519-530.

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics, 17*(9), 847-848.

Zhou, X., Guang, X., Sun, D., Xu, S., Li, M., Seim, I., . . . Yang, G. (2018). Population genomics of finless porpoises reveal an incipient cetacean species adapted to freshwater. *Nat Commun, 9*(1), 1276. doi:10.1038/s41467-018-03722-x