Title: High quality whole genome sequence of an abundant Holarctic odontocete, the harbour porpoise (*Phocoena phocoena*)

Authors: Marijke Autenrieth[1], Stefanie Hartmann[2], Ljerka Lah[1,3], Anna Roos[4], Alice B. Dennis[1], Ralph Tiedemann[1*]

[1] University of Potsdam, Institute of Biochemistry and Biology, Unit of Evolutionary Biology/Systematic Zoology, 14476 Potsdam, Germany

[2] University of Potsdam, Institute of Biochemistry and Biology, Evolutionary Adaptive Genomics, 14476 Potsdam, Germany

[3] current address: Novartis BTDM Mengeš, Kolodvorska 27, SI-1234 Mengeš, Slovenia (All work in regards to this manuscript was performed at University of Potsdam[1])

[4] Swedish Museum of Natural History, SE-104 05 Stockholm, Sweden

***Corresponding author**: Prof. Dr. Ralph Tiedemann tiedeman@uni-potsdam.de

Abstract

The harbour porpoise (*Phocoena phocoena*) is a highly mobile cetacean found in waters across the Northern hemisphere. It occurs in coastal water and inhabits water basins that vary broadly in salinity, temperature, and food availability. These diverse habitats could drive differentiation among populations. Here we report the first harbour porpoise genome, assembled *de novo* from a Swedish Kattegat individual. The genome is one of the most complete cetacean genomes currently available, with a total size of 2.7 Gb and 50% of the total length found in just 34 scaffolds. Using the largest 122 scaffolds, we were able to validate a high level of homology to the chromosome-level genome assembly of the closest related species for which such resource was available, the domestic cattle (*Bos taurus*). The draft annotation comprises 22,154 predicted gene models, which we further annotated through matches to the NCBI nucleotide database, GO categorization, and motif prediction. To infer the adaptive abilities of this species, as well as their population history, we performed a Bayesian skyline analysis, and produced results that are concordant with the demographic history of this species, including expansion and fragmentation events. Overall, this genome assembly, together with the draft annotation, represents a crucial addition to the limited genetic markers currently available for the study of porpoises and Phocoenidae conservation, phylogeny, and evolution.

Introduction

A central question in evolutionary biology is understanding the genomic basis of key innovations and adaptations, the respective impacts of selection *vs.* drift on specific genes, and how these patterns vary across the entire genome. In non-model organisms, these questions have been typically addressed using just a few genetic markers (e.g. microsatellites, single mitochondrial or nuclear genes), often in the absence of knowledge about the genomic position of the analyzed loci (e.g., Diefenbach et al., 2015; Ellegren & Galtier, 2016; Hoelzel et al., 1999; Lopes et al., 2015; Wiemann et al., 2010). However, in the last ~10 years the cost of sequencing has decreased, while read quality and sequence length have increased (Neely et al., 2018), and genome-wide resources are no longer limited to a few model organisms. With these resources, it is now possible to detect more subtle differences among species and populations and to extend these beyond nucleotide polymorphisms to larger structural variants, linkage groups, and highly divergent genes and regions on a genomic scale (e.g. Johannesson et al., 2017; L. F. Li et al., 2017; Whiteman, 2017; Willoughby et al., 2017).

Among the non-model organisms, cetaceans are a valuable group of animals to invest in creating solid genetic resources; cetaceans attract considerable attention, both scientifically and in the general public, and genomic resources enable further examination of their adaptive evolution and facilitate population assessments in the context of conservation and management. As fully aquatic mammals, cetaceans encounter escalating levels of mostly human-caused threats that lead to death, infertility, and reduction of population size and stability (Fietz et al., 2013; Godard-Codding et al., 2011). These include noise pollution (Dyndo et al., 2015; Nabe-Nielsen et al., 2014), marine debris and by-catch (Scheidat et al., 2008; Unger et al., 2017), and infectious diseases (Siebert et al., 2001; van Beurden et al., 2017).

Beside acoustic (Amundin, 2016) and photo-id monitoring, a number of genetic markers are frequently and intensively used to estimate population status and boundaries. The most commonly implemented markers are microsatellites (e.g., Fontaine et al., 2010; Hoffman & Amos, 2005; Palo et al., 2001) and mitochondrial DNA (e.g., Autenrieth et al., 2018; Graves et al., 2009; Speller et al., 2016), and more recently genome-wide single nucleotide polymorphism (SNP) surveys (e.g., Lah et al., 2016). New genomic resources hence do not only facilitate evolutionary research, but enable elevation of conservation genetics approaches to the genomic level (Foote et al., 2015).

Genetic resources available for cetaceans are growing, but do not yet adequately cover the diversity of this group (Foote et al., 2016; Keane et al., 2015; Nery et al., 2013; Sun et al., 2013; Yim et al., 2013; Zhou et al., 2013). Thus far sequenced genomes include three baleen whales (DeWoody et al., 2017; Kishida et al., 2015; Yim et al., 2013) and five species (with a collective seven genomes) of the toothed whales (Foote et al., 2015; Jones et al., 2017; Lindblad-Toh et al., 2011; Neely et al., 2018; Warren et al., 2017; Zhou et al., 2013), but the completeness of these assemblies varies (Table 1). Within the odontoceti only four of the ten families (*Physeteroidae*, *Lipotidae*, *Monodontidae* and *Delphinidae*) have been whole-genome sequenced to date. The family *Phocoenidae*, sister group to the *Monodontidae*, encompasses seven species, three of them threatened (Gatesy et al., 2013; Geisler et al., 2011, Rojas-Bracho et al. 2017, Wang et al., 2017). Being the smallest cetaceans in stature, phocoenids occur close to the coast and are therefore, like river dolphins, especially prone to human induced threats. This puts these species particularly at threat of decline or even extinction, as in the case of the Vaquita, endemic to Mexico (Taylor et al., 2017); yet no full genomic resource is available for these species.

We have chosen to develop a high quality genetic resource of the harbour porpoise (*Phocoena phocoena*). This cetacean species occurs from sub-polar to temperate coastal waters in the Northern hemisphere (Fontaine et al., 2017; Gaskin, 1984; Lah et al., 2016). As an apex predator, the harbour porpoise (*Phocoena phocoena*) is a key indicator for conservation and biodiversity measurements in the Nordic Seas (Hooker & Gerber, 2004; Lawrence et al., 2016; Sergio et al., 2008).

The population size of harbour porpoises in European Atlantic Shelf waters is estimated to be 375,000 with shifts across the last decade in the exact regions they occupy (e.g. in the North Sea; (Hammond et al., 2013)). Estimates of population size in the western Baltic Sea are smaller at approximately 40,000 animals (Benke et al., 2014; Scheidat et al., 2008; Viquerat et al., 2014), while the Baltic Sea proper only has an estimated < 500 individuals (Amundin, 2016), classifying this population as critically endangered (Benke et al., 2014; Hammond et al., 2008; Scheidat et al., 2008).

In addition to occupying different habitats, these populations differ in a number of features, including food preference (Hammond et al., 2013), activity patterns (Nuuttila et al., 2017), and fine scale morphological features. These differences are corroborated by genetic differentiation (Fontaine et al., 2012, 2014; Wiemann et al., 2010) and significant isolation by distance (Lah et al., 2016). This led to the suggestion that different ecotypes of harbour porpoise in the North Atlantic and Baltic Sea exist (Fontaine et al., 2014, 2017; Galatius et al., 2012). Given this evidence for genetic differentiation among populations, a genomic resource would not only be a valuable addition to the few existing cetacean genomes, but it would also be very useful for monitoring the separate porpoise populations and investigating whether there are specific local adaptations, e.g., to differences in temperature and salinity, in these small cetaceans.

We present here the first *de novo* assembly of the full genome of the harbor porpoise, scaffolded with *in vitro* proximity ligation data (hereafter "Chicago" library), and draft-annotated to predict its coding proteins and their functions (Deposited at NCBI as BioProject: PRJNA417595 with BioSample-ID: SAMN08000480). We demonstrate a high level of completeness of the assembly, showing that many of our scaffolds are near-chromosome-level, and provide insight into past demographic dynamics of the population the sequenced individual originated from, using a Bayesian skyline plot (Li & Durbin, 2011).

Materials and Methods

DNA sampling

Tissue for whole genome sequencing came from a single individual from the Kattegat sea (Glommen - Falkenberg), Sweden (ID: C2009/02665). Muscle tissue was sampled in July 2009 from a by-caught female of probably young age (22.4kg, 110.5m), frozen and transported to Potsdam, Germany for DNA extraction. Sample preparation and Genomic DNA isolation were performed following the Quiagen DNeasy Blood & Tissue Kit (Cat 69506, Hilden). Successful high molecular weight DNA-isolation was confirmed by Sanger sequencing of the mitochondrial control region and visualization of fragment sizes of the entire extraction using a Tape Station (Agilent 2200, Santa Clara, CA 95051). By mtDNA control region sequencing we verified that the analyzed specimen carried haplotype PHO7 (Tiedemann et al., 1996), indicative of the separate Beltsea population of the Kattegat/Western Baltic Sea region (Lah et al., 2016; Wiemann et al., 2010).

Genome sequencing and assembly

The draft *de novo* assembly was constructed from two libraries (insert sizes ca. 300 and ca. 500bp); sequenced in 125bp PE reads on the Illumina HiSeq 2500 at EUROFINS Genomics.

Between 90.43% and 91.52% of the bases had a quality of 30 or higher; the mean quality values ranged between 34.39 and 34.78. Reads were trimmed using CUTADAPT v1.10 (Martin, 2011), and an initial assembly was performed using SOAPDENOVO2 (Luo et al., 2015). DNA from the same sample was used by Dovetail Genomics for construction of the Chicago library (Putnam et al., 2016) and sequenced in 150bp PE reads on an Illumina NextSeq500 at the University of Potsdam. The draft assembly was then scaffolded with the Chicago library data using the HiRise pipeline (Dovetail Genomics LLC, Santa Cruz, CA, USA).

BLOBTOOLS was run to examine potential contaminants, based on divergence in GC-content and read coverage variation across the assembly (Laetsch & Blaxter, 2017). Presence of core, single copy, and orthologous genes was measured using CEGMA and BUSCO with the latter run in the genome mode for the Laurasiatheria database (Simão et al., 2015).

Genome annotation

Genome annotation was performed using MAKER2 (Holt & Yandell, 2011), which makes use of different programs and draws from several lines of evidence. Prior to annotation, repetitive elements were soft-masked with REPEATMASKER (Smit et al., 2013-2015) using the te_protein repeat database (Smith et al., 2007). In the first MAKER2 run, three gene predictors were used: SNAP (Bromberg et al., 2008) was *ab initio* trained with the CEGMA results (Parra et al., 2007), GENEMARK-ES (Ter-Hovhannisyan et al., 2008) was run using an HMM produced by *ab initio* training on the whole *P. phocoena* genome, AUGUSTUS was run using the presets for human, as is recommended for vertebrates (Stanke et al., 2004). Protein sequences, supplied as evidence, were obtained from the complete UniProt database (553,941 Proteins) plus NCBI entries of 184,527 proteins (retrieved on 20 March 2017) with matches to the following

keywords: "*Balaenopteridae*", "*Lipotes vexillifer*", "*Neophocaena*", "*Orcinus orca*", "*Phocoena*", "*Physeter catodon*", "*Pontoporia blainvillei*", "*Tursiops truncatus*".

For the second MAKER2 run, we created a new SNAP-HMM based on the first MAKER2 output and ran it with the same parameters as the first run, exchanging only the SNAP HMM and excluding the protein evidence. The resulting CDS predictions were extracted from the final gff file, which was created by *fathom* implemented in SNAP (Bromberg et al., 2008). These gene predictions were further verified by a BLASTN search against the entire GENBANK non-redundant (nr) nucleotide sequence database (date downloaded 21.07.2017) using a maximum number of 3 target sequences and an e-value of 1e-10. Summary statistics were generated using GENOME ANNOTATION GENERATOR (Hall et al., 2014). We then used all CDS and their BLAST results in BLAST2GO (Goetz et al., 2008) to identify conserved protein domains with INTERPROSCAN. We functionally annotated the CDS with GO terms, which are a controlled vocabulary to further categorize gene function (Ashburner et al., 2000; Carbon et al., 2017).

Comparative genomics

The closest relative to *P. phocoena* with a chromosome-level assembly currently available is the domestic cattle, *Bos taurus*. To validate our assembly, we compared our scaffolds to the *B. taurus* chromosomes (assembly UMD 3.1.1 downloaded from NCBI, ACCESSION DAAA00000000). Specifically, the 122 *P. phocoena* scaffolds of at least 1Mbp were aligned to the *B. taurus* chromosomes using the nucmer software of the MUMMER package v. 3.23 (Kurtz et al., 2004). From the coordinates of these alignments, runs of ten or more consecutive matches of each at least 250bp between a given *P. phocoena* scaffold and a *B. taurus* chromosome were extracted using custom perl scripts. Their start and end positions

were used to generate a CIRCOS plot (http://circos.ca/) that shows regions of collinearity as well as rearrangements. For the CIRCOS plot, separate ribbons are displayed between a *B. taurus* chromosome and a *P. phocoena* scaffold for consecutive hits that were each no more than 20,000 bp apart. If a hit was more than 20,000bp from the next run of consecutive hits, a new ribbon was started (Figure 1).

Demographic history

Using genome-wide diploid sequence data, it is possible to reconstruct the population history by estimating population sizes in the past (Li et al., 2011). To estimate the demographic history of the sequenced individual, phased sequence data is required. Since we did not have this, we generated the diploid sequence data as follows: we used the SNP frequency spectrum based on our genome assembly, which is a haploid sequence, and the PE reads used to construct the *de novo* assembly prior to Chicago scaffolding (described above, we used both insert sizes). These reads were first mapped back to the final assembly using BWA (Li & Durbin, 2009). SNP data was extracted from the resulting bam files, and variants were extracted using SAMTOOLS vs.1.6. (Li, Handsaker, et al., 2009) and BCFTOOLS (Li, Handsaker, et al., 2009). Results were then used to generate the final Bayesian skyline plot in the PSMC package, using perl scripts psmc2history.pl and psmc_plot.pl (Li et al., 2011) and including 100 bootstraps (Li et al., 2011). The parameters of the PSMC analysis were set following the recommendation from the authors (Li & Durbin, 2011, https://github.com/lh3/psmc) and we assumed a generation time of 10 years (Birkun Jr. & Frantzis, 2008) and a mutation rate of 2.2 $\times 10^{-9}$ year/site (Taylor et al., 2007).

Results

De novo assembly of the P. phocoena genome

Shotgun sequencing produced a total of 1,268M reads (Table 1); these were used to generate a draft assembly with 2.4M sequences (including contigs and scaffolds) and an N50 of 33.1kb. This assembly was combined with the Chicago library data (528M read) for final scaffolding by Dovetail Genomics (Putnam et al., 2016). The final HiRise assembly contains 2,025,248 scaffolds in total and including singletons, with 13,498 scaffolds bigger than 1kb (Table 3), had a total length of 2.7Gb (N50 of 23.8Mb), and was covered by 87x sequence data. The greatest improvements from the addition of the Chicago libraries was in building up the 34 longest scaffolds, which make up approximately half of the entire assembly (Table 3). The BUSCO and CEGMA analyses also suggest that we have largely reconstructed the entire genome, as we identified 96.9% (91.3% complete) of the 2,586 Eukaryotic and 94.2 (88.7% complete) of the 6,253 Laurasiatheria BUSCO core genes and 90% of the 248 ultra-conserved CEGMA genes (54% complete).

Genome completeness and annotation

The MAKER2 annotation resulted in a predicted 22,154 coding genes (Table 4). In total, 21,750 CDS had a BLAST hit against the nucleotide database, which accounts for 98% of the total CDSs. Of these BLAST hits, 99% were from vertebrates, and these were dominated (90%) by hits to cetacea (thereof 59% *Tursiops truncatus*, 27% *Orcinus orca*). Further annotation with INTERPROSCAN revealed 250,126 features of these predicted proteins. These comprise hits in several protein domain databases, e.g. 23,319 PFAM protein domains, 37,046 PANTHER gene families, 24,538 SUPERFAMILY annotations and 31,114 GENE3D domains.

Assignment of the BLAST results to Gene Ontology (GO) categories resulted in 55,143 hits across the GO categories (Figure 2).

Comparison to a chromosome-level assembly

To illustrate the completeness of the harbour porpoise genome, we mapped the 122 biggest scaffolds against the domestic cattle chromosomes, the closest available genome with chromosomal information. Between the 30 domestic cattle chromosomes and the 122 harbour porpoise scaffolds, a total of 24,394 separate matches (as described in the Materials and Methods section) were constructed. The matching regions show some rearrangement (Figure 1). In some cases, ribbons of different colors (coming from different domestic cattle chromosomes) match a single porpoise scaffold. For example, one porpoise scaffold (Sc_11228) encompasses parts of the cow chromosome 1 (violet), cow chromosome 8 (red) and cow chromosome number 27 (blue, Figure 1). The scaffold IDs that match each domestic cattle chromosome are listed in Supplemental Material 1. We find almost complete matching of cow chromosome to a single porpoise scaffold for cow chromosome 25/ porpoise scaffold Sc_12150 (cyan-ribbon). Additionally cow chromosomes 12, 23, 24, 26 and 27 are each matched by only two porpoise scaffolds. Other chromosomes are matched by three or more scaffolds, with a maximum of ten (x-chromosome; Supplement Material 1).

Inference of Kattegat/Baltic population history

We inferred the demographic history of the harbour porpoise *P. phocoena* Kattegat/Baltic population based on one single individual (Li et al., 2011) using the PSMC algorithm and all of our PE read data. In the resulting graph (Figure 3), the inferred population size (N_e) began to increase around 1.5 Myr ago, reaching an N_e of 500,000 during the following 1 Myr. The

estimated population size peaked approximately 200,000 years ago at an estimated N_e ~500,000, before it dropped to a fifth of the original size around 100,000 years ago, leading to a lower N_e of 100,000.

Discussion

We present here a high quality *de novo* genome assembly for the harbour porpoise *Phocoena phocoena*. Scaffolding with Chicago Library method (i.e. *in vitro* proximity ligation) has previously produced high quality genomes in plants (Moll et al., 2017; Paajanen et al., 2017), insects (Nosil et al., 2018; Wachi et al., 2017), and birds (Holt et al., 2018). We have applied this method to a sample from a deceased by-caught cetacean specimen, demonstrating that this approach can be successfully used for non-model mammal species and samples of lower quality, as is typical for cetacean research samples that are often obtained from stranded or by-caught individuals (e.g., Autenrieth et al., 2018; Ottewell et al., 2016; Sabatier et al., 2015; Squadrone et al., 2015).

With a GC-content of 41.4% and a total length of 2.7 GB, this assembly is comparable to other high quality cetacean genomes (Table 1; Groenen et al., 2013; Zimin et al., 2009). In support of this, the scaffold N50 and sequence coverage are similar to the recently published high quality genomes of the beluga whale (Jones et al., 2017), sperm whale (Warren et al., 2017), and bottlenose dolphin (Neely et al., 2018) (Table 1). Scans for core genes from the BUSCO and CEGMA databases identified a near completeness of these genes in the assembly and support that we have largely reconstructed the entire genome. Only 122 scaffolds are needed, including the 34 largest scaffolds representing 50% of the whole genome, to almost completely cover the chromosomes of the *B. taurus* genome (Figure 1). The CIRCOS plot depicting this (Figure 1) illustrates the near-completeness of the porpoise genome reached by

theses 122 longest scaffolds and shows large regions of colinearity with the chromosome-level *B. taurus* genome assembly. Of these largest scaffolds some likely represent entire chromosomes, indicated by single *B. taurus* chromosomes, which are largely matched by only one or two porpoise scaffolds (e.g., cow chromosome 25, see Supplementary Information 1). Other *B. taurus* chromosomes are in only 2-3 pieces among our scaffolds, e.g. cow chromosomes 12 and 24 (Supplementary Information 1). Some of these fragmented matches likely indicate chromosomal rearrangements between domestic cattle and the harbour porpoise, two species diverged approximately 60Myrs ago within the Cetartiodactyla (Gatesy et al., 2013). Chromosomal rearrangements are seen several times among distinct lineage of Cetartiodactyla (Avila et al., 2015; Kulemzina et al., 2009, 2011; Pauciullo et al., 2014), for example between camel, pig and domestic cattle (Balmus et al., 2007). Based on these comparisons, we infer that our assembly represents a nearly complete genome of *P. phocoena* and that our largest scaffolds are nearly-complete chromosomes.

The draft annotation using MAKER2 resulted in a total number of 22,154 annotated genes. This number is comparable to other published cetacean genomes, including 21,459 genes predicted in the bottlenose dolphin (Lindblad-Toh et al., 2011), 20,605 genes predicted in the minke whale (Yim et al., 2013), and 22,711 predicted genes in the grey whale (DeWoody et al., 2017). The annotated genes appear to broadly span key functional gene categories (biological processes, cellular components and molecular function), both across the annotated GO terms and the INTERPROSCAN results. Nonetheless this draft annotation should be further improved using transcriptome data (Fitak et al., 2016; Reyes-Chin-Wo et al., 2017) which were not available for our study organism. So far, with this draft annotation it is now possible to directly locate and further investigate key genes that may be under selection or underlie specific adaptive traits, both in cetaceans in general and in specific lineages (Chikina

et al., 2016; McGowen et al., 2014; Yim et al., 2013). This genomic resource also provides candidate genes potentially underlying the local adaptation previously postulated within the harbor porpoise (Fontaine et al., 2014, 2017; Galatius et al., 2012). This includes genes with functions associated with low salinity and other cetacean specific traits (e.g., osmoregulation (aquaporines), blubber (ELOVL), reduction of smell, and taste receptors (Kishida et al., 2015; Pedro et al., 2015; Wang et al., 2015; Xue et al., 2014)). Sequence data available for gene families of interest will allow for comparative analyses across the cetacean/odontocete lineages, with a focus on porpoises representing the smallest cetaceans.

This genomic resource will also enable future genome-level analyses in the Phocoenidae, especially work that uses the *P. phocoena* genome to aid in genome assembly and annotation in other Phocoenidae species. This assembly contributes to the growing lists of whale genomes, and is the first sequenced genome for this family, which exhibits different ecological and morphological adaptations to the marine environment (Racicot et al., 2016; Wisniewska et al., 2016).

Our *P. phocoena* genome assembly will facilitate population genomic analyses within this species. It enables whole genome resequencing of further individuals to improve or revise subspecies and population delimitation (Rosel et al., 1999; Fontaine et al., 2014) or to allow identification of populations threatened by low effective population size/genetic isolation (Gui et al., 2013; Nadachowska-Brzyska et al., 2016; Park et al., 2015; Shafer et al., 2017). For the harbour porpoise this genome will further enable mapping of so far anonymous nuclear microsatellite (Wiemann et al. 2010) and SNP (Lah et al. 2016) loci, thus facilitating genomic monitoring of critically endangered population. It may be used for the compilation of informative SNP panels for population assignment in regions of population admixture, like in the Baltic region (Wiemann et al. 2010; Lah et al. 2016).

A first step in utilizing the *P. phocoena* genome was to investigate the demographic history of the harbor porpoise using the pairwise sequentially Markovian coalescent (PSMC) model. The PSMC estimates the demographic history of one diploid individual using the distance between locally estimated times to the most recent common ancestor (Li et al., 2011; Schraiber & Akey, 2015). Comparing the coalescence of alleles across the genome at a specific time points can indicate population contraction (many regions) or expansion (fewer regions). However this interpretation is only valid when one large panmictic population is assumed (Mazet et al., 2016), and it has been shown that population structure biases estimations of N_e (Beichman et al., 2017; Li et al., 2011; Nadachowska-Brzyska et al., 2016). For example, the pattern of few large and many small pairwise differences could be observed both for populations that have experienced a bottlenecked, but also when a there is a multi-deem population structure (Nielsen & Beaumont, 2009). Following the recommendation of Li and Durbin (2011) to be cautious about inferences reaching further back in time than 1 Mio years, we concentrate here on the inference of potential population status between 20,000 and 1 Mio years ago (Nadachowska-Brzyska et al., 2016). Our sequenced specimen, sampled in the Baltic Sea, was assigned with high likelihood to the Beltsea population (Lah et al., 2016) by mtDNA analysis (exhibiting haplotype PHO 7; cf. Tiedemann et al., 1996; Wiemann et al., 2010). The Baltic Sea was formed only after the last glacial and – in its present form - is less than 10,000 years old (Andrén et al., 2011). Hence, for the time frame of our PSMC analysis, the sequenced specimen most likely represents the North Atlantic shelf population, a population now spanning from Iceland through the North Sea up to the Kattegat (Lah et al., 2016). Around 1 Mio years ago, our analysis infers an increase of N_e from around 200,000 to an estimated 500,000 around 400,000 years ago, which could indicate an expansion of the Atlantic Shelf population, assuming a panmictic population. During the last interglacial

period, the Eemian, the inferred N_e remained relatively high. The abrupt inferred N_e decline around 100kya ago, taken together with the formation of sea ice during the last glacial period could indicate a population bottleneck or a fragmentation into separate subpopulations (Nielsen et al., 2009). Among cetaceans (Brünchen-Olsen et al., n.d.; Foote et al., 2016; Yim et al., 2013; Zhou et al., 2013), the observed pattern, i.e., an increase in N_e followed by a sudden drop, is most similar to the demographic history of bottlenose dolphin (*Tursiops truncatus*), a related species with a similar North Atlantic distribution. However, one should treat these inferences on historical N_e with caution, as (1) they contain untested assumptions about panmixis, (2) other demographic scenarios could yield a similar pattern (Nielsen et al., 2009), and (3) the resolution of PSMC decreases for more recent times (Nadachowska-Brzyska et al, 2016).

In summary, we present here the first whole genome assembly and annotation of the harbour porpoise, at this point the most complete assembly for the Family *Phocoenidae*. This genome adds to available Cetacean genomes by supplying important resources for further investigation within the *Odontoceti* and *Cetacea*. This will provide an invaluable resource for further genetic studies of the harbour porpoise, both for whole-genome investigations into population structure and to identify key genes associated with local adaptation. This genome represents a crucial genetic resource for further investigation in the conservation, population genetics and phylogeny of other *Phocoenidae*, including the currently most rare marine mammal, the almost extinct Vaquita (*Phocoena sinus*) (Taylor et al., 2017).

Acknowledgments

Financial support came from the Bundesamt für Naturschutz (FKZ # 3514824600), as part of a larger study of population genomics. We thank Prof. Dr. Michael Hofreiter for providing access to the Illumina NextSeq platform. Additional support came from the University of Potsdam. Large-scale computations were carried out on resources provided by Prof. Dr. Michael Lenhard at University of Potsdam and the High Performance Computing Cluster Orson2, managed by ZIM (Zentrum für Informationstechnologie und Medienmanagement) at the University of Potsdam.

References

Amundin, M. (2016). SAMBAH - Static Acoustic Monitoring of the Baltic Sea Harbour porpoise. LIFE Project Number LIFE 08 NAT/S/000261 European Commission, 77pp. available at http://www.sambah.org/SAMBAH-Final-Report-FINAL-for-website-April-2017.pdf.

Andrén, T., Björck, S., Andrén, E., Conley, D., Zillén, L., & Anjar, J. (2011). The Baltic Sea Basin, 75–97. doi:10.1007/978-3-642-17220-5

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene ontologie: Tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. doi:10.1038/75556

Autenrieth, M., Ernst, A., Deaville, R., Demaret, F., IJsseldijk, L. L., Siebert, U., & Tiedemann, R. (2018). Putative origin and maternal relatedness of male sperm whales (*Physeter macrocephalus*) recently stranded in the North Sea. *Mammalian Biology*, *88*, 156–160. doi:10.1016/j.mambio.2017.09.003

Avila, F., Baily, M. P., Merriwether, D. A., Trifonov, V. A., Rubes, J., Kutzler, M. A., … Raudsepp, T. (2015). A cytogenetic and comparative map of camelid chromosome 36 and the minute in alpacas. *Chromosome Research*, *23*(2), 237–251. doi:10.1007/s10577-014-9463-3

Balmus, G., Trifonov, V. A., Biltueva, L. S., O'Brien, P. C. M., Alkalaeva, E. S., Fu, B., … Ferguson-Smith, M. A. (2007). Cross-species chromosome painting among camel, cattle, pig and human: Further insights into the putative Cetartiodactyla ancestral karyotype. *Chromosome Research*, *15*(4), 499–515. doi:10.1007/s10577-007-1154-x

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

Benke, H., Bräger, S., Dähne, M., Gallus, A., Hansen, S., Honnef, C. G., … Verfuß, U. K. (2014). Baltic Sea harbour porpoise populations: Status and conservation needs derived from recent survey results. *Marine Ecology Progress Series*, *495*, 275–290. doi:10.3354/meps10538

Birkun Jr., A. A., & Frantzis, A. (2008). *Phocoena phocoena ssp. relicta.* The IUCN Red List

of Threatened Species 2008: e.T17030A6737111. Downloaded on 12 October 2017.

Bromberg, Y., Yachdav, G., & Rost, B. (2008). SNAP predicts effect of mutations on protein function. *Bioinformatics*, *24*(20), 2397–2398. doi:10.1093/bioinformatics/btn435

Brünchen-Olsen, A., Westerman, R., Kazmierczyk, Z., Vertyankin, V. V., Godard-Codding, C., Bickham, J. W., & DeWoody, J. A. (n.d.). The inference of gray whale (*Eschrichtius robustus*) population attributes from whole-genome sequences.

Carbon, S., Dietze, H., Lewis, S. E., Mungall, C. J., Munoz-Torres, M. C., Basu, S., … Westerfield, M. (2017). Expansion of the gene ontology knowledgebase and resources: The gene ontology consortium. *Nucleic Acids Research*, *45*(D1), D331–D338. doi:10.1093/nar/gkw1108

Chikina, M., Robinson, J. D., & Clark, N. L. (2016). Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Molecular Biology and Evolution*, *33*(9), 2182–2192. doi:10.1093/molbev/msw112

DeWoody, J. A., Fernandez, N. B., Brüniche-Olsen, A., Antonides, J. D., Doyle, J. M., San Miguel, P., … Bickham, J. (2017). Characterization of the gray whale (*Eschrichtius robustus*) genome and a genotyping array based on single nucleotide polymorphisms in candidate genes. *Biological Bulletin*, *232*(June), 186–197.

Dyndo, M., Wiśniewska, D. M., Rojano-Doñate, L., & Madsen, P. T. (2015). Harbour porpoises react to low levels of high frequency vessel noise. *Scientific Reports*, *5*, 11083. doi:10.1038/srep11083

Fietz, K., Graves, J. A., & Olsen, M. T. (2013). Control Control Control: A reassessment and comparison of GenBank and chromatogram mtDNA sequence variation in Baltic Grey Seals (*Halichoerus grypus*). *PLoS ONE*, *8*(8), 1–7. doi:10.1371/journal.pone.0072853

Fitak, R. R., Mohandesan, E., Corander, J., & Burger, P. A. (2016). The de novo genome assembly and annotation of a female domestic dromedary of North African origin. *Molecular Ecology Resources*, *16*(1), 314–324. doi:10.1111/1755-0998.12443

Fontaine, M. C., Roland, K., Calves, I., Austerlitz, F., Palstra, F. P., Tolley, K. A., … Aguilar, A. (2014). Postglacial climate changes and rise of three ecotypes of harbour porpoises, *Phocoena phocoena*, in western Palearctic waters. *Molecular Ecology*, *23*(13), 3306–3321. doi:10.1111/mec.12817

Fontaine, M. C., Snirc, A., Frantzis, A., Koutrakis, E., Oztürk, B., Oztürk, A. a, & Austerlitz, F. (2012). History of expansion and anthropogenic collapse in a top marine predator of the Black Sea estimated from genetic data. *Proceedings of the National Academy of Sciences of the USA*, *109*(38), E2569-76. doi:10.1073/pnas.1201258109

Fontaine, M. C., Thatcher, O., Ray, N., Piry, S., Brownlow, A., Davison, N. J., … Goodman, S. J. (2017). Mixing of porpoise ecotypes in southwestern UK waters revealed by genetic profiling. *Royal Society Open Science*, *4*(160992), 1–14. doi:10.1098/rsos.160992

Fontaine, M. C., Tolley, K. A., Michaux, J. R., Birkun, A., Ferreira, M., Jauniaux, T., … Baird, S. J. E. (2010). Genetic and historic evidence for climate-driven population fragmentation in a top cetacean predator: the harbour porpoises in European water. *Proceedings. Biological Sciences / The Royal Society*, *277*, 2829–2837. doi:10.1098/rspb.2010.0412

Foote, A. D., Liu, Y., Thomas, G. W. C., Vinař, T., Alföldi, J., Deng, J., … Gibbs, R. A. (2015). Convergent evolution of the genomes of marine mammals. *Nature Genetics*, *47*(3), 272–5. doi:10.1038/ng.3198

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., … Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, *7*(May), 11693. doi:10.1038/ncomms11693

Galatius, A., Kinze, C. C., & Teilmann, J. (2012). Population structure of harbour porpoises in the Baltic region: evidence of separation based on geometric morphometric comparisons. *Journal of the Marine Biological Association of the United Kingdom*, *92*(8), 1669–1676. doi:10.1017/S0025315412000513

Gaskin, D. (1984). The harbour porpoise *Phocoena phocoena (L.)*: regional populations, status, and information on direct and indirect catches. *Reports of the International Whaling Commission*, *34*, 569–586.

Gatesy, J., Geisler, J. H., Chang, J., Buell, C., Berta, A., Meredith, R. W., … McGowen, M. R. (2013). A phylogenetic blueprint for a modern whale. *Molecular Phylogenetics and Evolution*, *66*(2), 479–506. doi:10.1016/j.ympev.2012.10.012

Geisler, J. H., McGowen, M. R., Yang, G., & Gatesy, J. (2011). A supermatrix analysis of genomic, morphological, and paleontological data from crown Cetacea. *BMC Evolutionary Biology*, *11*(1), 112. doi:10.1186/1471-2148-11-112

Godard-Codding, C. A. J., Clark, R., Fossi, M. C., Marsili, L., Maltese, S., West, A. G., … Stegeman, J. J. (2011). Pacific ocean-wide profile of CYP1A1 expression, stable carbon and nitrogen isotope ratios, and organic contaminant burden in sperm whale skin biopsies. *Environmental Health Perspectives*, *119*(3), 337–343. doi:10.1289/ehp.0901809

Goetz, S., Garccia-Gomez, M. J., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., … Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research*, *36*(10), 3420–3435. doi:10.1093/nar/gkn176

Graves, J. A., Helyar, A., Biuw, M., Jüssi, M., Jüssi, I., & Karlsson, O. (2009). Microsatellite and mtDNA analysis of the population structure of grey seals (*Halichoerus grypus*) from three breeding areas in the Baltic Sea. *Conservation Genetics*, *10*(1), 59–68. doi:10.1007/s10592-008-9517-1

Groenen, M. A. M., Archibald, A. L., Uenishi, H., Tuggle, C. K., Takeuchi, Y., Rothschild, M. F., … Hunt, T. (2013). Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, *491*(7424), 393–398. doi:10.1038/nature11622

Gui, D., Jia, K., Xia, J., Yang, L., Chen, J., Wu, Y., & Yi, M. (2013). De novo assembly of the Indo-Pacific Humpback Dolphin leucocyte transcriptome to identify putative genes involved in the aquatic adaptation and immune response. *PLoS ONE*, *8*(8). doi:10.1371/journal.pone.0072417

Hall, B., DeRego, T., & Geib, S. (2014). GAG: the Genome Annotation Generator (Version 1.0) [Software]. Available from http://genomeannotation.github.io/GAG.

Hammond, P. S., Bearzi, G., Bjørge, A., Forney, K. A., Karczmarski, L., Kasuya, T., … Wilson, B. (2008). *Phocoena phocoena* (Baltic Sea subpopulation). (errata version published in 2016) The IUCN Red List of Threatened Species 2008: e.T17031A98831650. Downloaded on 10 October 2017. Retrieved from http://www.iucnredlist.org/details/17031/0

Hammond, P. S., Macleod, K., Berggren, P., Borchers, D. L., Burt, L., Canadas, A., … Vazquez, J. A. (2013). Cetacean abundance and distribution in European Atlantic shelf waters to inform conservation and management. *Biological Conservation*, *164*, 107–122. doi:10.1016/j.biocon.2013.04.010

Hoffman, J. I., & Amos, W. (2005). Microsatellite genotyping errors: Detection approaches, common sources and consequences for paternal exclusion. *Molecular Ecology*, *14*(2), 599–612. doi:10.1111/j.1365-294X.2004.02419.x

Holt, C., Campbell, M., Keays, D. A., Edelman, N., Kapusta, A., Maclary, E., … Shapiro, M. D. (2018). Improved Genome Assembly and Annotation for the Rock Pigeon (*Columba*

livia). *G3: Genes|Genomes|Genetics, early onli*. doi:10.1534/g3.117.300443

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1), 491. doi:10.1186/1471-2105-12-491

Hooker, S. K., & Gerber, L. (2004). Marine reserves as a tool for ecosystem-based management: the potential importance of megafauna. *BioScience*, *54*(1), 27–39. doi:10.1641/0006-3568(2004)054[0027:MRAATF]2.0.CO;2

Jones, S., Taylor, G., Chan, S., Warren, R., Hammond, S., Bilobram, S., … Haulena, M. (2017). The genome of the Beluga Whale (*Delphinapterus leucas*). *Genes*, *8*(12), 378. doi:10.3390/genes8120378

Keane, M., Semeiks, J., Webb, A. E., Li, Y. I., Quesada, V., Craig, T., … deMagalhães, J. P. (2015). Insights into the evolution of longevity from the bowhead whale genome. *Cell Reports*, *10*(1), 112–122. doi:10.1016/j.celrep.2014.12.008

Kishida, T., Thewissen, J., Hayakawa, T., Imai, H., & Agata, K. (2015). Aquatic adaptation and the evolution of smell and taste in whales. *Zoological Letters*, *1*(1), 9. doi:10.1186/s40851-014-0002-z

Kulemzina, A. I., Trifonov, V. A., Perelman, P. L., Rubtsova, N. V., Volobuev, V., Ferguson-Smith, M. A., … Graphodatsky, A. S. (2009). Cross-species chromosome painting in Cetartiodactyla: Reconstructing the karyotype evolution in key phylogenetic lineages. *Chromosome Research*, *17*(3), 419–436. doi:10.1007/s10577-009-9032-3

Kulemzina, A. I., Yang, F., Trifonov, V. A., Ryder, O. A., Ferguson-Smith, M. A., & Graphodatsky, A. S. (2011). Chromosome painting in Tragulidae facilitates the reconstruction of Ruminantia ancestral karyotype. *Chromosome Research*, *19*(4), 531–539. doi:10.1007/s10577-011-9201-z

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

Laetsch, D. R., & Blaxter, M. L. (2017). BlobTools: Interrogation of genome assemblies. *F1000Research*, *6*(1287), 1–15. doi:10.12688/f1000research.12232.1 and doi:10.5281/zenodo.845347

Lah, L., Trense, D., Benke, H., Berggren, P., Gunnlaugsson, Þ., Lockyer, C., … Tiedemann, R. (2016). Spatially explicit analysis of genome-wide SNPs detects subtle population structure in a mobile marine mammal, the Harbor Porpoise. *PLoS ONE*, *11*(10), 1–23. doi:10.1371/journal.pone.0162792

Lawrence, J. M., Armstrong, E., Gordon, J., Lusseau, S. M., & Fernandes, P. G. (2016). Passive and active, predator and prey: using acoustics to study interactions between cetaceans and forage fis. *ICES Journal of Marine Science*, *73*(8), 2075–2084. doi:10.1093/icesjms/fsw013

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

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. doi:10.1093/bioinformatics/btp352

Lindblad-Toh, K., Garber, M., Zuk, O., Lin, M. F., Parker, B. J., Washietl, S., … Kellis, M. (2011). A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, *478*(7370), 476–482. doi:10.1038/nature10530

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … Wang, J. (2015). Erratum: SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, *4*(1), 30. doi:10.1186/s13742-015-0069-2

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10. doi:10.14806/ej.17.1.200

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

McGowen, M. R., Gatesy, J., & Wildman, D. E. (2014). Molecular evolution tracks macroevolutionary transitions in Cetacea. *Trends in Ecology and Evolution*. Elsevier Ltd. doi:10.1016/j.tree.2014.04.001

Moll, K. M., Zhou, P., Ramaraj, T., Fajardo, D., Devitt, N. P., Sadowsky, M. J., … Mudge, J. (2017). Strategies for optimizing BioNano and Dovetail explored through a second reference quality assembly for the legume model, *Medicago truncatula*. *BMC Genomics*, *18*(1), 1–16. doi:10.1186/s12864-017-3971-4

Nabe-Nielsen, J., Sibly, R. M., Tougaard, J., Teilmann, J., & Sveegaard, S. (2014). Effects of noise and by-catch on a Danish harbour porpoise population. *Ecological Modelling*, *272*, 242–251. doi:10.1016/j.ecolmodel.2013.09.025

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Molecular Ecology*, *25*(5), 1058–1072. doi:10.1111/mec.13540

Neely, B. A., Ellisor, D. L., & Davis, W. C. (2018). Proteomics as a metrological tool to evaluate genome annotation accuracy following de novo genome assembly: a case study using the Atlantic bottlenose dolphin (*Tursiops truncatus*). *bioRxiv*, 254250. doi:10.1101/254250

Nery, M. F., Gonzalez, D. J., & Opazo, J. C. (2013). How to make a dolphin: molecular signature of positive selection in cetacean genome. *PLoS ONE*, *8*(6), 2–8. doi:10.1371/journal.pone.0065491

Nielsen, R., & Beaumont, M. A. (2009). Statistical inferences in phylogeography. *Molecular Ecology*, *18*(6), 1034–1047. doi:10.1111/j.1365-294X.2008.04059.x

Nosil, P., Villoutreix, R., Carvalho, C. F. de, Farkas, T. E., Soria-Carrasco, V., Feder, J. L., … Gompert, Z. (2018). Natural selection and the predictability of evolution in Timema stick insects. *Science*, *359*(6377), 765–770. doi:10.1126/SCIENCE.AAP9125

Nuuttila, H. K., Courtene-Jones, W., Baulch, S., Simon, M., & Evans, P. G. H. (2017). Don't forget the porpoise: acoustic monitoring reveals fine scale temporal variation between bottlenose dolphin and harbour porpoise in Cardigan Bay SAC. *Marine Biology*, *164*(3), 1–16. doi:10.1007/s00227-017-3081-5

Paajanen, P., Kettleborough, G., Lopez-Girona, E., Giolai, M., Heavens, D., Baker, D., … Clark, M. D. (2017). A critical comparison of technologies for a plant genome sequencing project. *bioRxiv*, 201830. doi:10.1101/201830

Palo, J. U., Makinen, H. S., Helle, E., Stenman, O., & Vainola, R. (2001). Microsatellite variation in ringed seals (*Phoca hispida*): genetic structure and history of the Baltic Sea population. *Heredity*, *86*, 609–617.

Park, J. Y., An, Y. R., Kanda, N., An, C. M., An, H. S., Kang, J. H., … Kim, H. (2015). Cetaceans evolution: Insights from the genome sequences of common minke whales. *BMC Genomics*, *16*(1), 13. doi:10.1186/s12864-015-1213-1

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core

genes in eukaryotic genomes. *Bioinformatics*, *23*(9), 1061–1067. doi:10.1093/bioinformatics/btm071

Pauciullo, A., Perucatti, A., Cosenza, G., Iannuzzi, A., Incarnato, D., Genualdo, V., … Iannuzzi, L. (2014). Sequential cross-species chromosome painting among river buffalo, cattle, sheep and goat: A useful tool for chromosome abnormalities diagnosis within the family bovidae. *PLoS ONE*, *9*(10). doi:10.1371/journal.pone.0110297

Pedro, S. L. S., Alves, J. M. P., Barreto, A. S., & De Souza Lima, A. O. (2015). Evidence of positive selection of aquaporins genes from *pontoporia blainvillei* during the evolutionary process of cetaceans. *PLoS ONE*, *10*(7), 1–11. doi:10.1371/journal.pone.0134516

Putnam, N. H., Connell, B. O., Stites, J. C., Rice, B. J., Hartley, P. D., Sugnet, C. W., … Rokhsar, D. S. (2016). Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Research*, *26*, 342–350. doi:10.1101/gr.193474.115

Racicot, R. A., Gearty, W., Kohno, N., & Flynn, J. J. (2016). Comparative anatomy of the bony labyrinth of extant and extinct porpoises (Cetacea: Phocoenidae). *Biological Journal of the Linnean Society*, *119*(4), 831–846. doi:10.1111/bij.12857

Reyes-Chin-Wo, S., Wang, Z., Yang, X., Kozik, A., Arikit, S., Song, C., … Michelmore, R. W. (2017). Genome assembly with in vitro proximity ligation data and whole-genome triplication in lettuce. *Nature Communications*, *8*, 14953. doi:10.1038/ncomms14953

Scheidat, M., Gilles, A., Kock, K. H., & Siebert, U. (2008). Harbour porpoise *Phocoena phocoena* abundance in the southwestern Baltic Sea. *Endangered Species Research*, *5*(2–3), 215–223. doi:10.3354/esr00161

Schraiber, J. G., & Akey, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, *16*(12), 727–740. doi:10.1038/nrg4005

Sergio, F., Caro, T., Brown, D., Clucas, B., Hunter, J., Ketchum, J., … Hiraldo, F. (2008). Top Predators as Conservation Tools: Ecological Rationale, Assumptions, and Efficacy. *Annual Review of Ecology, Evolution, and Systematics*, *39*(1), 1–19. doi:10.1146/annurev.ecolsys.39.110707.173545

Shafer, A. B. A., Peart, C. R., Tusso, S., Maayan, I., Brelsford, A., Wheat, C. W., & Wolf, J. B. W. (2017). Bioinformatic processing of RAD-seq data dramatically impacts downstream population genetic inference. *Methods in Ecology and Evolution*, *8*(8), 907–917. doi:10.1111/2041-210X.12700

Siebert, U., Wünschmann, A., Weiss, R., Frank, H., Benke, H., & Frese, K. (2001). Post-mortem findings in harbour porpoises (*phocoena phocoena*) from the German North and Baltic Seas. *Journal of Comparative Pathology*, *124*(2–3), 102–114. doi:10.1053/jcpa.2000.0436

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. doi:10.1093/bioinformatics/btv351

Smit, A. F. A., Hubley, R., & Green, P. (n.d.). RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

Smith, C. D., Edgar, R. C., Yandell, M. D., Smith, D. R., Celniker, S. E., Myers, E. W., & Karpen, G. H. (2007). Improved repeat identification and masking in Dipterans. *Gene*, *389*(1), 1–9. doi:10.1016/j.gene.2006.09.011

Speller, C., van den Hurk, Y., Charpentier, A., Rodrigues, A., Gardeisen, A., Wilkens, B., … Hofreiter, M. (2016). Barcoding the largest animals on Earth: ongoing challenges and molecular solutions in the taxonomic identification of ancient cetaceans. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150332.

doi:10.1098/rstb.2015.0332
Stanke, M., Steinkamp, R., Waack, S., & Morgenstern, B. (2004). AUGUSTUS: A web server for gene finding in eukaryotes. *Nucleic Acids Research*, *32*(Web Server issue), W309–W312. doi:10.1093/nar/gkh379
Sun, Y. B., Zhou, W. P., Liu, H. Q., Irwin, D. M., Shen, Y. Y., & Zhang, Y. P. (2013). Genome-wide scans for candidate genes involved in the aquatic adaptation of dolphins. *Genome Biology and Evolution*, *5*(1), 130–139. doi:10.1093/gbe/evs123
Taylor, B. L., Chivers, S. J., Larese, J., & Perrin, W. F. (2007). Generation length and percent mature estimates for IUCN assessments of cetaceans. *Administrative Report LJ-07-01 National Marine Fisheries*, 24. doi:10.1.1.530.4789
Taylor, B. L., Rojas-Bracho, L., Moore, J., Jaramillo-Legorreta, A., Ver Hoef, J. M., Cardenas-Hinojosa, G., … Hammond, P. S. (2017). Extinction is imminent for Mexico's endemic porpoise unless fishery bycatch is eliminated. *Conservation Letters*, *10*(5), 588–595. doi:10.1111/conl.12331
Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., & Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Cold Spring Harbor Laboratory Press*, *18*, 1979–1990. doi:10.1101/gr.081612.108
Tiedemann, R., Harder, J., Gmeiner, C., & Haase, E. (1996). Mitochondrial DNA sequence patterns of Harbour porpoises (*Phocoena phocoena*) from the North and the Baltic Sea. *Zeitschrift Für Säugetierkunde*, *61*, 104–111.
Unger, B., Herr, H., Benke, H., Böhmert, M., Burkhardt-Holm, P., Dähne, M., … Siebert, U. (2017). Marine debris in harbour porpoises and seals from German waters. *Marine Environmental Research*, 1–8. doi:10.1016/j.marenvres.2017.07.009
van Beurden, S. J., Ijsseldijk, L. L., van de Bildt, M. W. G., Begeman, L., Wellehan, J. F. X., Waltzek, T. B., … Penzes, J. J. (2017). A novel cetacean adenovirus in stranded harbour porpoises from the North Sea: detection and molecular characterization. *Archives of Virology*, *162*(7), 2035–2040. doi:10.1007/s00705-017-3310-8
Viquerat, S., Herr, H., Gilles, A., Peschko, V., Siebert, U., Sveegaard, S., & Teilmann, J. (2014). Abundance of harbour porpoises (*Phocoena phocoena*) in the western Baltic, Belt Seas and Kattegat. *Marine Biology*, *161*(4), 745–754. doi:10.1007/s00227-013-2374-6
Wachi, N., Matsubayashi, K. W., & Maeto, K. (2017). Application of next-generation sequencing to the study of non-model insects. *Entomological Science*, (2017), 3–11. doi:10.1111/ens.12281
Wang, J., Yu, X., Hu, B., Zheng, J., Xiao, W., Hao, Y., … Wang, D. (2015). Physicochemical evolution and molecular adaptation of the cetacean osmoregulation-related Gene UT-A2 and Implications for Functional Studies. *Scientific Reports*, *5*(8795), 1–17. doi:10.1038/srep08795
Warren, W. C., Kuderna, L., Alexander, A., Catchen, J., Pérez-Silva, J. G., López-Otín, C., … Wise John Pierce, S. (2017). The novel evolution of the sperm whale genome. *Genome Biology and Evolution*, *9*(12), 3260–3264. doi:10.1093/gbe/evx187
Wiemann, A., Andersen, L. W., Berggren, P., Siebert, U., Benke, H., Teilmann, J., … Tiedemann, R. (2010). Mitochondrial Control Region and microsatellite analyses on harbour porpoise (*Phocoena phocoena*) unravel population differentiation in the Baltic Sea and adjacent waters. *Conservation Genetics*, *11*(1), 195–211. doi:10.1007/s10592-009-0023-x
Wisniewska, D. M. M., Johnson, M., Teilmann, J., Rojano-Doñate, L., Shearer, J., Sveegaard,

S., … Madsen, P. T. T. (2016). Ultra-High Foraging Rates of Harbor Porpoises Make Them Vulnerable to Anthropogenic Disturbance. *Current Biology*, *26*(11), 1441–1446. doi:10.1016/j.cub.2016.03.069

Xue, X., Feng, C. Y., Hixson, S. M., Johnstone, K., Anderson, D. M., Parrish, C. C., & Rise, M. L. (2014). Characterization of the fatty acyl elongase (elovl) gene family, and hepatic elovl and delta-6 fatty acyl desaturase transcript expression and fatty acid responses to diets containing camelina oil in Atlantic cod (*Gadus morhua*). *Comparative Biochemistry and Physiology Part - B: Biochemistry and Molecular Biology*, *175*(1), 9–22. doi:10.1016/j.cbpb.2014.06.005

Yim, H.-S., Cho, Y. S., Guang, X., Kang, S. G., Jeong, J.-Y., Cha, S.-S., … Lee, J.-H. (2013). Minke whale genome and aquatic adaptation in cetaceans. *Nature Genetics*, *46*(1), 88–92. doi:10.1038/ng.2835

Zhou, X., Sun, F., Xu, S., Fan, G., Zhu, K., Liu, X., … Yang, G. (2013). Baiji genomes reveal low genetic variability and new insights into secondary aquatic adaptations. *Nature Communications*, *4*(2708), 1–6. doi:10.1038/ncomms3708

Zimin, A. V, Delcher, A. L., Florea, L., Kelley, D. R., Schatz, M. C., Puiu, D., … Salzberg, S. L. (2009). A whole-genome assembly of the domestic cow, *Bos taurus*. *Genome Biology*, *10*(4), R42. doi:10.1186/gb-2009-10-4-r42

Data Accessibility

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession PKGA00000000. The version described in this paper is version PKGA01000000. The draft annotation is deposited to Dryad (doi:10.5061/dryad.vr021gq).

Authors Contributions

R.T. and L.L. designed the study; A.R. provided the sample and associated biological information. L.L. performed molecular lab work, S.H. performed initial *de novo* assembly, M.A. executed all genome annotations and analyses, M.A., S.H., A.B.D., and R.T. analyzed and interpreted the results, M.A. wrote the draft manuscript. All authors edited and approved the final manuscript.

Tables and Figures

Table 1 Publicly available cetacean genomes in comparison to the new genome of the harbour porpoise (*Phocoena phocoena*)

Species name	NCBI-ID	WGS project	total sequence length in Gb	# of scaffolds	scaffolds N50	Genome coverage	publication
Phocoena phocoena	Pho_pho_vs1_UP	tba	2.68	13,498	23,753,446	87.0x	this study
Delphinapterus leucas	ASM228892v2	NQVZ01	2.36	6,972	19,885,328	117.0x	Jones et al. (2017)
Tursiops truncatus	NIST Tur_tru v1	MRVK01	2.13	2,648	26,555,543	114.5x	Neely et al. (2018)
Tursiops truncatus	Ttru_1.4	ABRN02	2.55	240,558	116,287	30x	Lindblad-Toh et al. (2011)
Orcinus orca	Oorc_1.1	ANOL02	2.37	1,668	12,735,091	200.0x	Foote et al. (2015)
Lipotes vexillifer	Lipotes_vexillifer_v1	AUPI01	2.43	30,713	2,419,148	115x	Zhou et al. (2013)
Physeter macrocephalus	Physeter_macrocephalus-2.0.2	AWZP01	2.28	11,711	427,290	75x	Warren et al (2017)
Physeter macrocephalus	ASM283717v1	PGGR01	2.52	59,237	37,183,148	248x	ds (Fan et al. 2017)
Eschrichtius robustus	ASM218922v1	NIPP01	2.85	57,203	187,455	11.0x	DeWoody et al. (2017)
Balaenoptera acutorostrata	BalAcu1.0	ATDI01	2.43	10,776	12,843,668	92x	Yim et al (2014)
Balaenoptera bonaerensis	ASM97880v1	BAUQ01	2.23	421,444	20,082	60x	Kishida et al. (2015)

NCBI-ID: assembly name, WGS project: whole genome sequence project, publications: ds = direct submission, not yet linked to a publication.

Table 2 Summary statistics of sequence data used for the draft and final HiRise assemblies.

	Insert (bp)	n reads	Coverage
Illumina Library 300	300	794 M	74.1 X
Illumina Library 500	500	473 M	44.2 X
Chicago library	1-50kb	528 M	87.2 X

Table 3 Assembly statistics of the harbour porpoise genome assembly.

	SOAPdenovo assembly	HiRise Assembly
Total length	2,669.6 Mb	2,681.2 Mb
Scaffolds		
N50 (number/length)	23,685bp / 0.032Mb	23,753,446bp/ 23.8Mb
N90 (number)	159,889 bp	43,146 bp
Longest	304,733 bp	67,078,619 bp
Number	2,139,681 bp	2,025,248 bp

Table 4 Genome annotation statistics

	Exons	Introns	Genes	CDS
Number	171,735	149,581	22,154	22,154
Longest (in bp)	10 kb	1,532 kb	2,299 kb	21 kb
Mean length	166	3,966	28,051	1,282
% genome covered by	-	-	23.2%	1.1%
GC% in CDS	-	-	-	54.48%

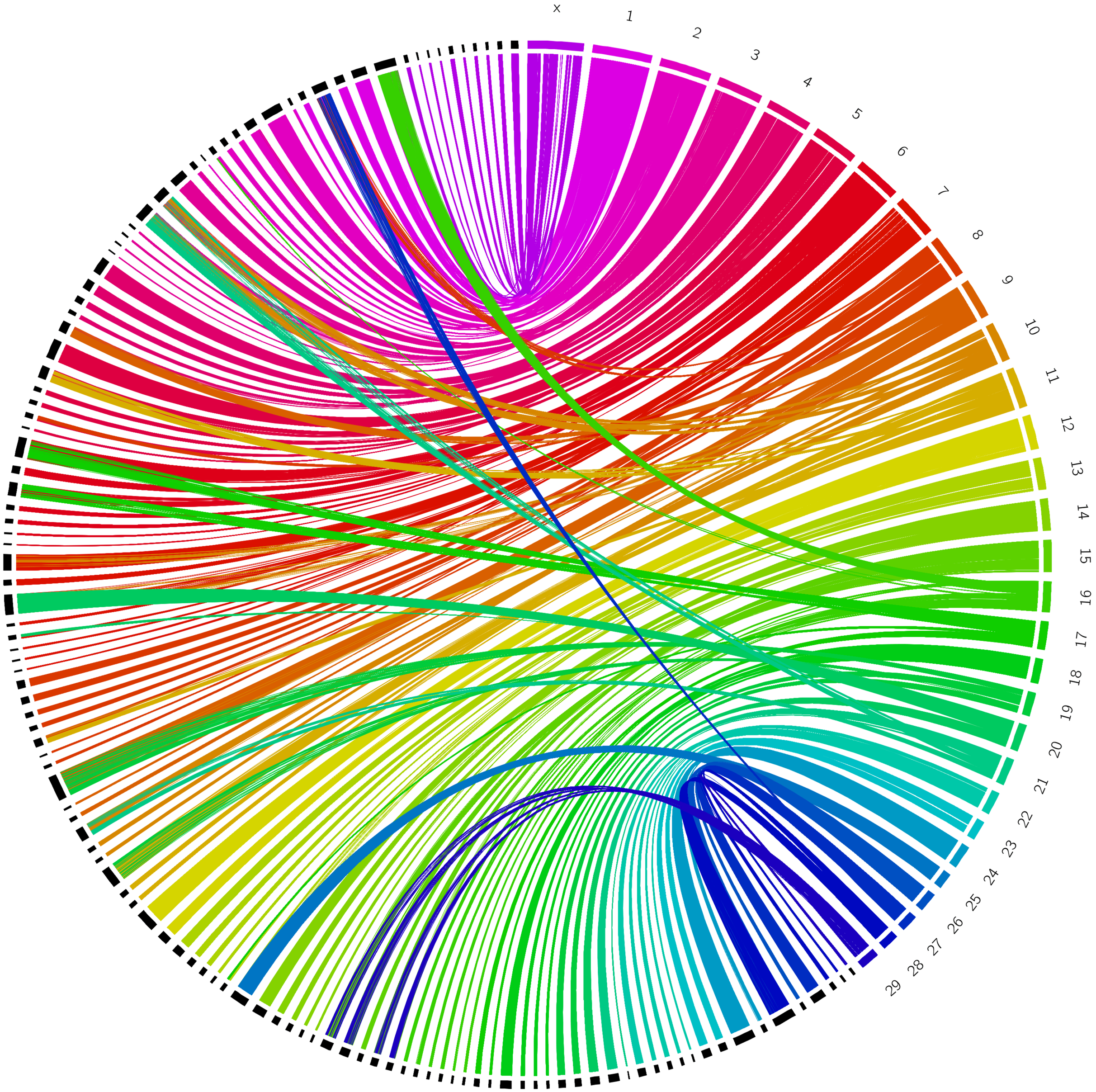
Figure *1* Comparison *of P. phocoena* scaffolds and *B. taurus* chromosomes. Each *B. taurus* nuclear chromosome is shown in a different colour on the right outer rim; they are labelled X for the X chromosome and 1-29 for the 29 autosomes. *P. phocoena* scaffolds of at least 1 Mbp are shown in black on the left outer rim and not labeled. Only the 122 largest scaffolds with consecutive MUMMER hits (as described in the main text) are included. Matches between *B. taurus* chromosomes and *P. phocoena* scaffolds are shown in the colour of the *B. taurus* chromosomes.

Figure 2 GO-Terms; GO annotation level 2, separated by gene ontology terms: "Cellular Component", "Biological Process", and "Molecular Function". Separate categories are listed, with the number of hits in parentheses.

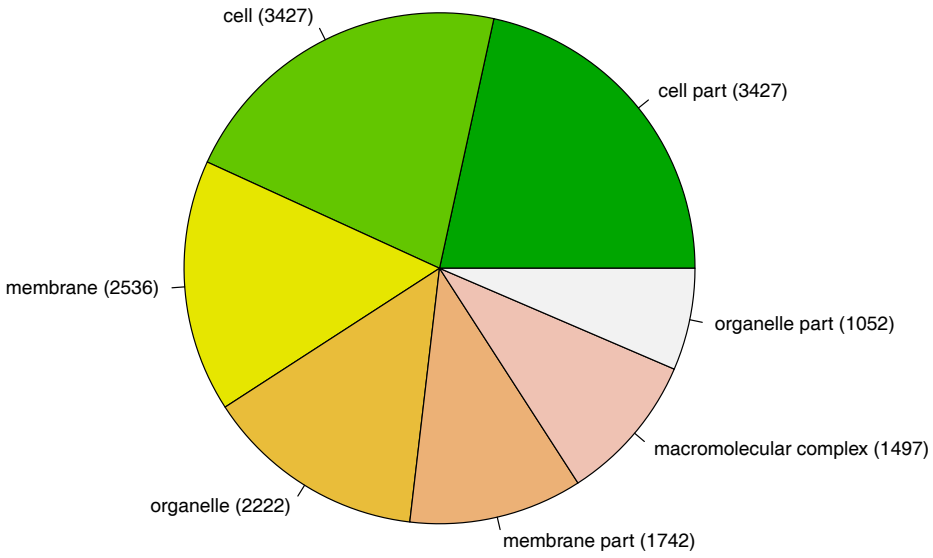
Figure 3 PSMC estimated harbour porpoise population size changes over time for the Baltic Sea. g = generation time; μ = mutation rate (per site, per year). Porpoise data generated on the basis of mapping PE reads to whole genome scaffolds during SNP calling. Pale lines show the trajectories of 100 bootstrap replicates.

Supplement Material 1 122 *Phocoena phocoena* scaffolds mapped in the CIRCOS plot, shown sorted to the *Bos taurus* chromosomes they match

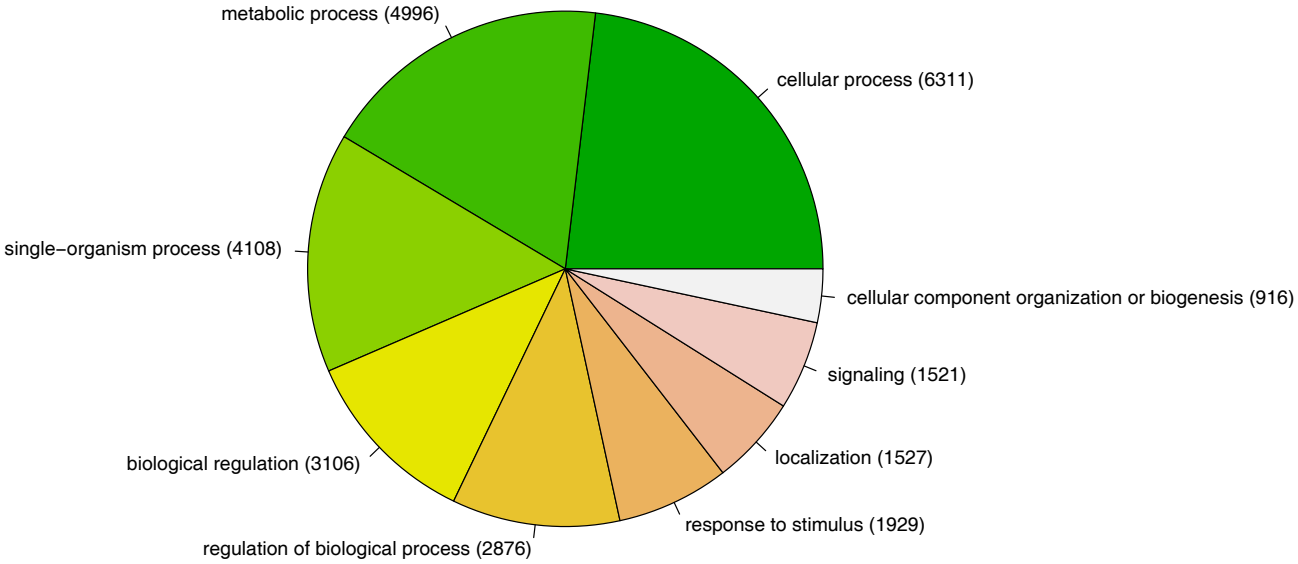
Bos taurus chromosomes	*Number of scaffolds matched*	*Phocoena phocoena* scaffold IDs
bt1	6	Sc_11228, Sc_12227, Sc_1990, Sc_5358, Sc_9797, Sc_9900
bt2	8	Sc_11626, Sc_12227, Sc_13447, Sc_529, Sc_6515, Sc_6894, Sc_7661, Sc_9857
bt3	9	Sc_11018, Sc_1195, Sc_12865, Sc_3364, Sc_6515, Sc_7569, Sc_7648, Sc_7661, Sc_9681
bt4	5	Sc_12949, Sc_1454, Sc_2312, Sc_2531, Sc_6983
bt5	6	Sc_10530, Sc_10690, Sc_1183, Sc_4380, Sc_8015, Sc_8313
bt6	7	Sc_10623, Sc_10772, Sc_11919, Sc_12799, Sc_3795, Sc_4462, Sc_8081
bt7	8	Sc_1186, Sc_12016, Sc_12231, Sc_13164, Sc_2192, Sc_3821, Sc_7425, Sc_8211
bt8	9	Sc_11228, Sc_1608, Sc_3409, Sc_4380, Sc_4484, Sc_4874, Sc_515, Sc_664, Sc_880
bt9	4	Sc_10089, Sc_12949, Sc_9322, Sc_9860
bt10	8	Sc_10140, Sc_1186, Sc_12231, Sc_12865, Sc_2049, Sc_3337, Sc_7569, Sc_9860
bt11	5	Sc_10530, Sc_12437, Sc_12947, Sc_1608, Sc_7385
bt12	2	Sc_1065, Sc_11493
bt13	6	Sc_10356, Sc_12150, Sc_4275, Sc_8289, Sc_9915, Sc_9960
bt14	5	Sc_1753, Sc_3476, Sc_4138, Sc_7403, Sc_958
bt15	5	Sc_10137, Sc_10729, Sc_5614, Sc_6829, Sc_7410
bt16	8	Sc_11693, Sc_12029, Sc_141, Sc_2264, Sc_529, Sc_5715, Sc_8603, Sc_9900
bt17	5	Sc_10623, Sc_4462, Sc_5093, Sc_8289, Sc_8998
bt18	4	Sc_1099, Sc_3393, Sc_7023, Sc_9717
bt19	4	Sc_2056, Sc_7385, Sc_8323, Sc_9860
bt20	3	Sc_12231, Sc_7191, Sc_8211
bt21	5	Sc_12865, Sc_3249, Sc_3337, Sc_7569, Sc_9533
bt22	5	Sc_10627, Sc_12496, Sc_13313, Sc_4865, Sc_77
bt23	2	Sc_5169, Sc_8212
bt24	2	Sc_3090, Sc_3285
bt25	1	Sc_12150
bt26	2	Sc_7704, Sc_982
bt27	2	Sc_11228, Sc_11897
bt28	3	Sc_2616, Sc_9108, Sc_982
bt29	5	Sc_10137, Sc_10729, Sc_2203, Sc_5614, Sc_6829,
btx	10	Sc_10048, Sc_12502, Sc_2350, Sc_4277, Sc_5542, Sc_5764, Sc_6046, Sc_610, Sc_6806, Sc_933


X
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29

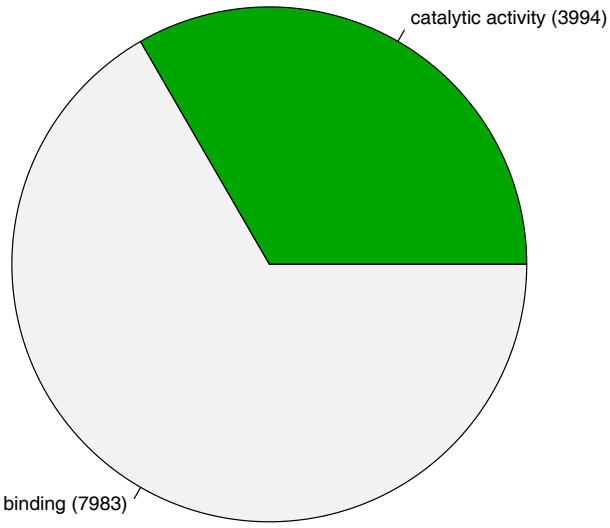
Cellular Component

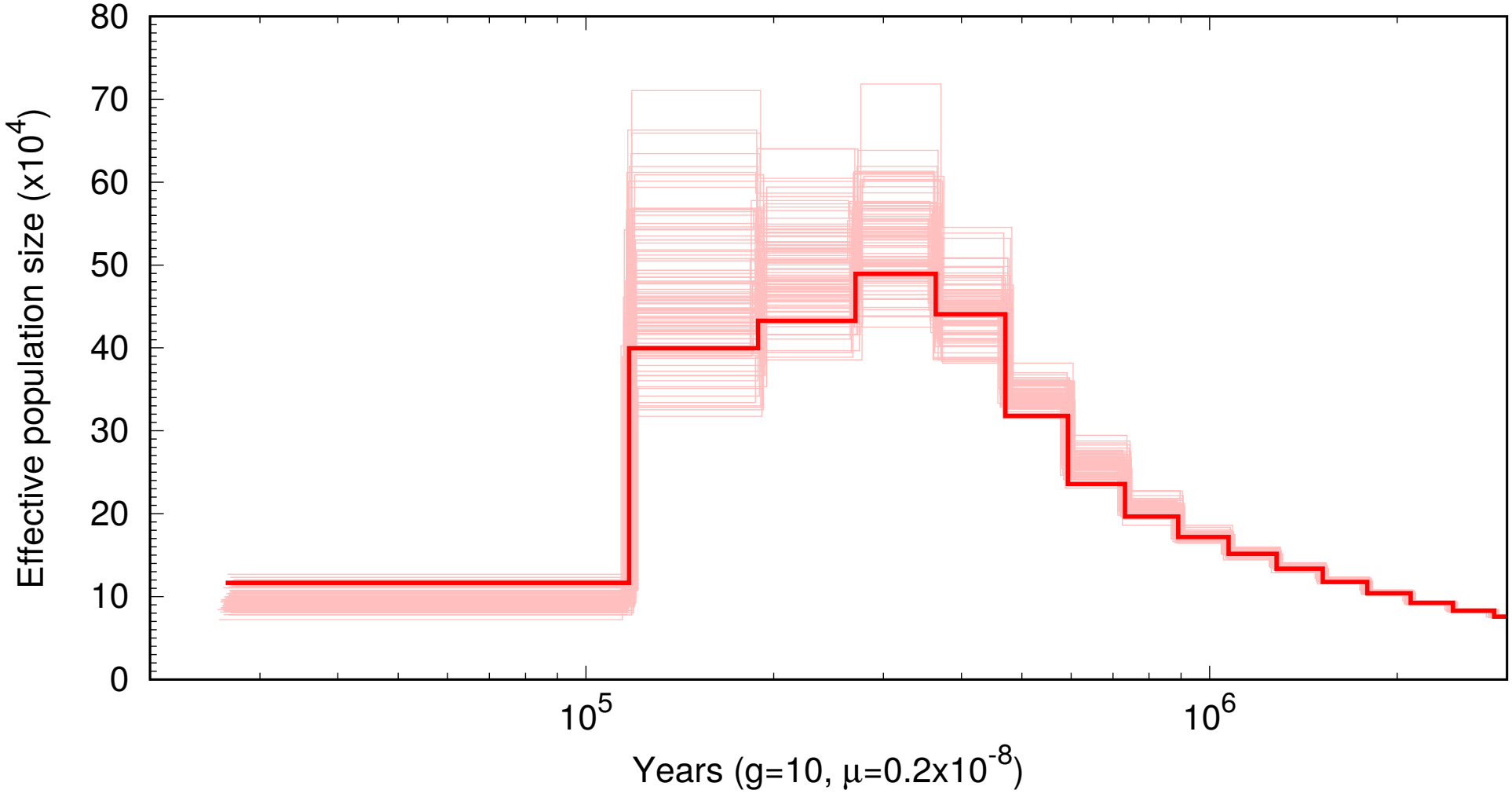


Biological Process



Molecular Function




80
70
60
50
40
30
20
10
0
Effective population size (x10^4)
10^5
10^6
Years (g=10, μ=0.2x10^-8)