EVOLUTIONARY BIOLOGY

Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow

Úlfur Árnason,[1]* Fritjof Lammers,[2,3,4]* Vikas Kumar,[2] Maria A. Nilsson,[2] Axel Janke[2,3,4]†

Reconstructing the evolution of baleen whales (Mysticeti) has been problematic because morphological and genetic analyses have produced different scenarios. This might be caused by genomic admixture that may have taken place among some rorquals. We present the genomes of six whales, including the blue whale (*Balaenoptera musculus*), to reconstruct a species tree of baleen whales and to identify phylogenetic conflicts. Evolutionary multilocus analyses of 34,192 genome fragments reveal a fast radiation of rorquals at 10.5 to 7.5 million years ago coinciding with oceanic circulation shifts. The evolutionarily enigmatic gray whale (*Eschrichtius robustus*) is placed among rorquals, and the blue whale genome shows a high degree of heterozygosity. The nearly equal frequency of conflicting gene trees suggests that speciation of rorqual evolution occurred under gene flow, which is best depicted by evolutionary networks. Especially in marine environments, sympatric speciation might be common; our results raise questions about how genetic divergence can be established.

INTRODUCTION

Baleen whales (Mysticeti) are strikingly derived marine mammals that encompass the largest animals living on Earth (*1*); however, their evolution is only poorly understood. Today, 15 species of extant baleen whales are known, and the fossil record includes many additional extinct species (*2*). The gigantic blue whale (*Balaenoptera musculus*) with a length of 30 m and a weight exceeding 150 metric tons and the fin whale (*Balaenoptera physalus*) are the largest animals on Earth (*1*). Both belong to the rorqual family (Balaenopteridae). Baleen whales have undergone significant adaptations to marine life and are characterized by their lack of teeth, which have been replaced by keratin bristles, the baleen that is used for filter feeding (*3*). It has been estimated that the blue whale takes in up to 3.6 metric tons of krill every day to supply the energy demand of their huge body sizes (*3*). The large body size of whales allowed them to occupy novel ecological niches by enabling deep dives and to endure long periods of starvation to reach feeding grounds (*4*). The evolutionary history of baleen whales is debated, despite extensive analyses of molecular and morphological characteristics (*2*, *5*). Moreover, molecular analyses of baleen whale evolution disagree with each other depending on the applied marker and type of phylogenetic analysis (*5*–*8*). Of particular interest are the humpback whales (*Megaptera novaeangliae*) and gray whales (*Eschrichtius robustus*), which are each placed in a separate genus or even in its own family, mainly based on analyses of their derived anatomy (*1*). However, these classifications are not supported by recent molecular studies, which suggest that they evolved from within rorquals, making the latter paraphyletic. To reflect this discordance, we will use the family name Balaenopteridae sensu lato, that is, including Balaenopteridae and Eschrichtiidae.

It is difficult to envision that the baleen whales evolved by allopatric speciation under vicariance because the marine environment largely lacks physical barriers for mobile species like whales (*1*, *9*). The study of the evolution of whales is further complicated by the fact that whales can hybridize. In the case of the blue whale and the fin whale, genetic analyses have shown that the female hybrid carried a fetus and had mated with a blue whale (*10*). Thus, these two species, as well as other rorquals, may not be entirely reproductively isolated. In addition, rorquals have a conserved karyotype of $2n = 44$ chromosomes and an identical chromosomal C-banding pattern, which facilitate producing fertile offspring (*11*).

Genomic analyses allow detailed insight into evolutionary processes such as speciation or past hybridization events (*12*) and permit examination of long-standing evolutionary questions (*13*). Introgressive hybridization, speciation with gene flow, and incomplete lineage sorting (ILS) may cause different local genealogies across the genome that can be detected by analyzing whole-genome sequences (*14*). Compared to terrestrial species, genomic data are limited for marine mammals, and before this study, genomic data were only available for three baleen whales: the bowhead whale (*Balaena mysticetus*), the minke whale (*Balaenoptera acutorostrata*), and the fin whale (*15*, *16*).

Here, we present genomic data of six mysticete species including the humpback and gray whale and the largest extant animal ever lived, the blue whale. The data are analyzed under the multispecies coalescent (MSC) that incorporates the genome-wide heterogeneity of gene trees to accurately infer speciation history (*14*). In addition, the genomes allow us to study signals of recent and ancestral introgression, to place divergences into a solid temporal context, and to explore genetic diversity and past demographic history of baleen whales.

RESULTS

Genome sequencing and assembly

Genomic DNA from six baleen whales and a hippopotamus (*Hippopotamus amphibius*) were sequenced with Illumina technology. Reference genome mapping of the whale genome data against the bowhead whale genome (*16*) yielded genome coverages of 6.3 to 27.2× (table S1). RepeatMasker (*17*) identified 40.3% repetitive sequences in the bowhead whale genome assembly. Of these, 6 and 18% were short and long interspersed elements (SINE and LINEs), respectively (table S2). Except for the genomic fraction of SINEs, these results are consistent with the original analyses of Keane *et al.* (*16*). We identified, on average, 25 million fixed single-nucleotide differences relative to the bowhead whale genome (table S3). Consensus sequences of all baleen whale genomes were aligned per scaffold, and repetitive sequences, gaps, and ambiguous bases were

[1]Department of Brain Surgery, Faculty of Medicine, University of Lund, Lund, Sweden. [2]Senckenberg Biodiversity and Climate Research Centre, Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt am Main, Germany. [3]Goethe University Frankfurt, Institute for Ecology, Evolution and Diversity, Biologicum, Max-von-Laue-Straße 13, 60439 Frankfurt am Main, Germany. [4]LOEWE Centre for Translational Biodiversity Genomics, Senckenberganlage 25, 60325 Frankfurt, Germany.
*These authors contributed equally to this work.
†Corresponding author. Email: axel.janke@senckenberg.de

removed. Empirical analyses and simulations using the approximate unbiased (AU) test (*18*) showed that 20–kilo–base pair (kbp) genome sequence alignments contain sufficient information for statistically significant maximum likelihood (ML) gene tree inference (figs. S1 to S3). The aligned scaffolds yielded 34,192 genome fragments (GFs), each 20 kbp long, totaling 643,840 kbp for each whale. This represents 49% of the nonrepetitive genome sequence. Sequencing the hippopotamus genome yielded 1,684,446,285 filtered reads and a sequencing depth of 55× (table S4). The reads were assembled de novo with Minia (*19*) and scaffolded with SSPACE, resulting in a genome assembly of 2.43 Gbp with a scaffold N50 of 120 kbp. AUGUSTUS (*20*) identified 29,998 coding sequences (CDSs); 37.0% of the genome were masked as repetitive (table S5).

The evolution of whales

Model testing identified the generalized time-reversible model with gamma-distributed rate variation with invariable sites (GTR + 4G + I) as the best-fitting nucleotide substitution model for the ML analyses of GFs. An MSC species tree of baleen whales based on 34,192 GF trees was supported with posterior probabilities of 1.0 for all branches (Fig. 1A and fig. S4). The topology conforms to previous nuclear gene and mitochondrial DNA (mtDNA) analyses (*5*, *21*) and a Bayesian phylogeny of the mtDNA sequences reported herein (fig. S5). The primary characteristic of the tree is the clear distinction between the Balaenidae (right whales) and the branch harboring the five rorquals plus the gray whale (Balaenopteridae sensu lato). The humpback whale (genus *Megaptera*) groups within the rorquals, resulting in a paraphyly of the current genus *Balaenoptera*. The gray whale of the monotypic family Eschrichtiidae is placed inside rorquals as a sister lineage to fin and humpback whale. However, quartet scores, that is, the support for any of three possible phylogenetic arrangements around an internal branch, identified conflict in resolving the branch leading to the ancestor of the gray, fin, and humpback whale (Fig. 1A, branch no. 3). The three possible topologies for this branch receive similar quartet scores (Fig. 1B), contrasting to a posterior probability of 1.0. Thus, we find highly similar support for placing the gray whale as a sister group to blue whales and sei whales or as a distinct clade outside the blue/sei/fin/humpback whale cluster. Somewhat inconclusive support also marks the first branch inside rorquals (Fig. 1B, branch no. 2) that places the minke whale as a sister lineage to the remaining Balaenopteridae sensu lato with a quartet score of 0.7. Phylogenetic conflict is also present in a CONSENSE (*22*) analysis of the GF trees. Although a majority-rule consensus tree confirms the coalescent-based species tree (Fig. 1A and fig. S6), two alternative phylogenetic positions of the gray whale are equally strongly represented (table S6).

The position of the gray whale in the species tree is supported by 10,315 (30.2%) GF trees compared to 8918 (26.1%) and 8721 (25.6%) GF trees, which place the gray whale in different positions inside rorquals.

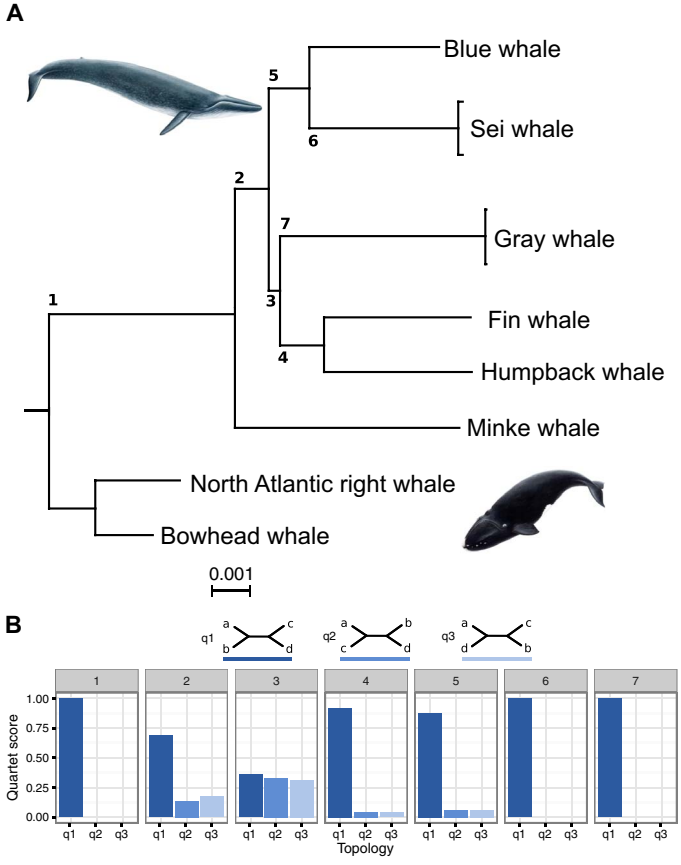

Fig. 1. MSC tree. (**A**) An MSC species tree was constructed from 34,192 individual GFs. Internal branches within Balaenopteridae are numbered 1 to 7. All branches receive maximal support ($P = 1.0$, ASTRAL analysis). Branch lengths were calculated from an ML analysis. Gray whales, family Eschrichtiidae, are placed inside Balaenopteridae as a sister group to fin and humpback whales. (**B**) ASTRAL quartet-score analyses for branches 1 to 7 (A). Quartet scores were calculated for the three possible arrangements (q1 to q3) for the respective branch. The principal quartet trees are depicted, with q1 representing the species tree. Branch nos. 2 and 3 receive only limited quartet scores, and no quartet can be significantly rejected.

A placement of the gray whale outside rorquals is supported by 3507 GF trees (10.3%). A consensus network analysis (*23*) of the GF trees yields a large cuboid structure of connecting alternative branches in the center of the network that indicates conflicting signals for the position of the gray whale inside rorquals (Fig. 2). At a threshold for conflicting edges of 11%, the grouping of the humpback and fin whale, the sei and blue whale, and the bowhead and North Atlantic right whale is unambiguous. At lower thresholds, the phylogenetic signal becomes more complex, indicating additional phylogenetic conflict in the data (fig. S7).

Gene flow analyses

D statistic (*24*) and D_{FOIL} (*25*) analyses identified several gene flow signals among rorquals (Fig. 3A and data S1 and S2). We find significant gene flow signals between minke whale and the ancestors of the blue and sei whale and those of the fin and humpback whale, respectively. The D_{FOIL} analyses find a strong signal for gene flow between the ancestor of the blue and sei whale and the ancestor of the fin and humpback whale, which is likely a phylogenetic signal related to a placement of placing the gray whale into different positions (Fig. 3A and data S1 and S2). In addition, signal for recent gene flow was inferred reciprocally from the blue whale to the fin and humpback whale for about 1 to 1.5% of the genome. The *D* statistic analyses also identified numerous signals for gene flow between the ancestor of the blue/sei whale and gray whale and that of the humpback whale and gray whale. Note that the *D* statistic and D_{FOIL} analyses depend on the species tree as in Fig. 1A and the signal may vary for other constellations. Our interpretation, therefore, focuses on signals that are independent of the evolutionary placement of gray whales.

In addition to character-based parsimony analysis, gene flow may preferably be studied by topology-based ML analysis using PhyloNet (*26*). PhyloNet identifies a statistically significant signal for gene flow between the minke whale and the ancestor of the other rorquals (Fig. 3B). With equal likelihood probability, gene flow occurred from the ancestor of the humpback and fin whale to that of the minke whale (Fig. 3C). Furthermore, with a topology change of the gray whale as a sister group to blue and sei whale, gene flow occurs from the ancestor of the blue and sei whale to that of the minke whale (Fig. 3D). Each of the three reticulations shows inheritance probabilities of about 33%, resembling the quartet-score distribution of the coalescent tree analyses (Fig. 1B).

Genetic diversity and population size history

Genome-wide heterozygosity varies considerably among baleen whales (Fig. 4A and fig. S8). At approximately 5×10^{-4} heterozygous sites per nucleotide, estimates were lowest for the gray whale, the minke whale, and the two sei whales. The blue whale genome shows the highest

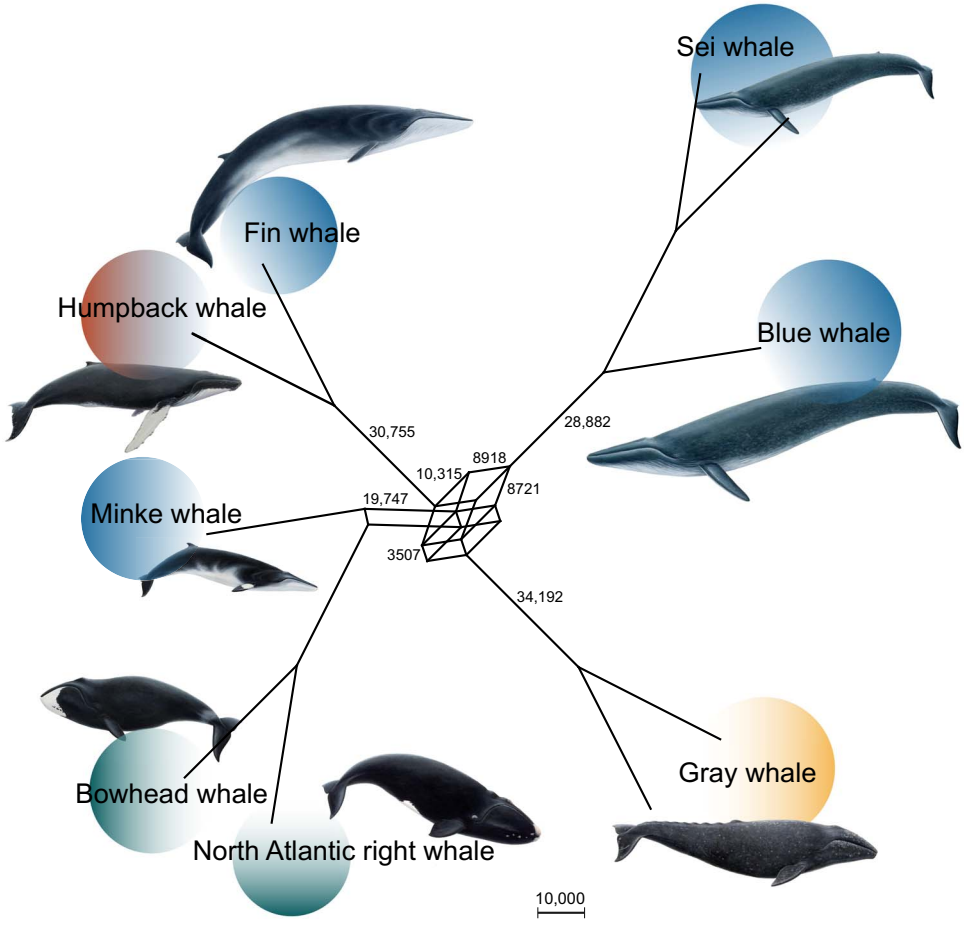

Fig. 2. Median network of 34,192 GF ML trees with 11% threshold. Conflicting evolutionary signals characterize the center of the network, which is equivalent to branch no. 3 in the species tree (Fig. 1). In addition, placing the minke whale has some conflicting signal, but the elongated rectangle indicates a higher degree of resolution. The number of supporting GFs is shown for selected splits. Colored circles indicate taxonomic classification. Blue, *Balaenoptera*; red, *Megaptera*; yellow, *Eschrichtius*; green, *Balaena* and *Eubalaena*.

degree of heterozygosity, which is elevated even when compared to other mammals (*27*). Estimates for heterozygosity in downsampled genomic data of blue whale were similar, minimizing the effects of potential artifacts by higher sequence coverage (fig. S9). The history of the effective population size (N_e) over the last 5 million years (Ma) was modeled from the distribution of heterozygous sites across the genome using a pairwise sequentially Markovian coalescent (PSMC) (*28*) analysis (Fig. 4B and fig. S10). Ancestral effective population sizes for all baleen whales, particularly the large blue, fin, and humpback whales, were notably higher during the Plio-Pleistocene transition (PPT; 2.6 Ma ago) than recent estimates (Fig. 4B). After the mid-Pleistocene transition (MPT), N_e of most baleen whales was relatively stable, until approximately 100 thousand years (ka) ago, the time of the last interglacial. After this time, baleen whale populations decreased. In contrast, gray whale population size remained stable during the interglacial, and its population size even increased in more recent times. The blue whale maintained a larger population size than other whales, but their numbers decreased at 400 ka ago after the MPT. The minke and fin whale population increased somewhat at 200 to 300 ka ago, followed by a steady decline. The N_e of the humpback whale was rather constant since 1 Ma ago and then shows a decline by two-thirds of its population at some 30 ka ago. Our estimates of historical population sizes of the fin and minke whale are consistent with previous analyses (*15*).

Divergence time estimates

The phylogenomic reconstruction of a paraphyletic position of Cetacea among Artiodactyla and the placement of the Hippopotamidae are, for the first time, supported by genomic sequence data analyses (Fig. 5). The divergence times are based on five calibration points (table S8). Hippopotamidae diverged at 53.5 Ma ago, close to the appearance of archaeocetes in the fossil record at 50 Ma ago (*29*). Rorquals diverged in the late Miocene, between 10.48 and 4.98 Ma ago (table S9). The divergence time between baleen and toothed whales at 30.5 Ma ago coincides with the Eocene/Oligocene transition at 33 Ma ago (*30*), which probably triggered the radiation of modern whales.

DISCUSSION

Our genome analyses have shown that the evolution of Balaenopteridae sensu lato (hereafter referred to as rorquals) is not characterized by an ordered dichotomous divergence of lineages as would be expected with respect to speciation in most other mammals. Coalescent-based analyses of more than 600-Mbp genomic data and network analyses show that the genomes of rorquals are characterized by contradicting genealogies for their central divergence. Thus, the evolution of rorquals appears to be a process of gradual divergences that likely gave rise to three lineages almost simultaneously: (i) blue plus sei whales, (ii) gray

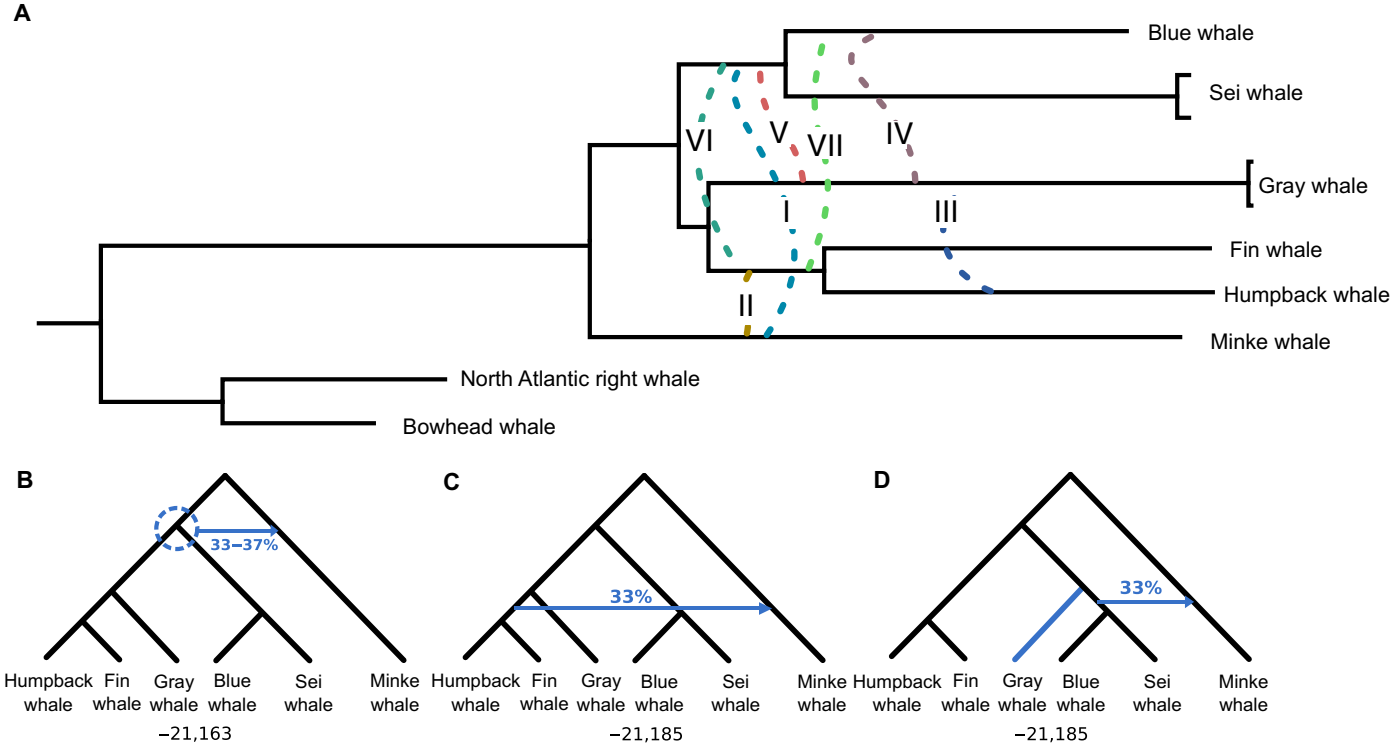

Fig. 3. Gene flow signals for baleen whales inferred by the *D* statistic, D_{FOIL}, and PhyloNet. (**A**) The species tree of baleen whales with gene flow signals detected by the *D* statistic and D_{FOIL} indicated by dashed lines. Signals I to IV were inferred by the *D* statistic, and signals V, VI, and VII were detected by D_{FOIL} and were partially corroborated by the *D* statistic. Note that D_{FOIL} cannot infer gene flow involving the minke whale. (**B** to **D**) Rooted networks for the Balaenopteridae sensu lato phylogeny with reticulations inferred from PhyloNet based on 34,192 20-kbp GFs. Reticulations are shown as blue arrows with inheritance probability denoted above or below. Log-likelihood scores are shown below the networks. Notably, inheritance probability around 33% resembles the distribution of quartet scores and the phylogenetic signals from GFs (Fig. 1). (B) The three best networks indicated a reticulation originating at the circled three branches to minke whale. Similar likelihood scores do not allow the identification of a single origin of gene flow; therefore, the networks were merged, and a range of inheritance probabilities is given. (C) The fourth best network has only a marginally poorer likelihood score and indicates a reticulation between the ancestor of the fin and humpback whale and that of the minke whale. (D) The fifth best network has the same likelihood as (C) and finds an alternative placement of gray whale (blue branch) and reticulation from the ancestor of the blue and sei whale to that of the minke whale.

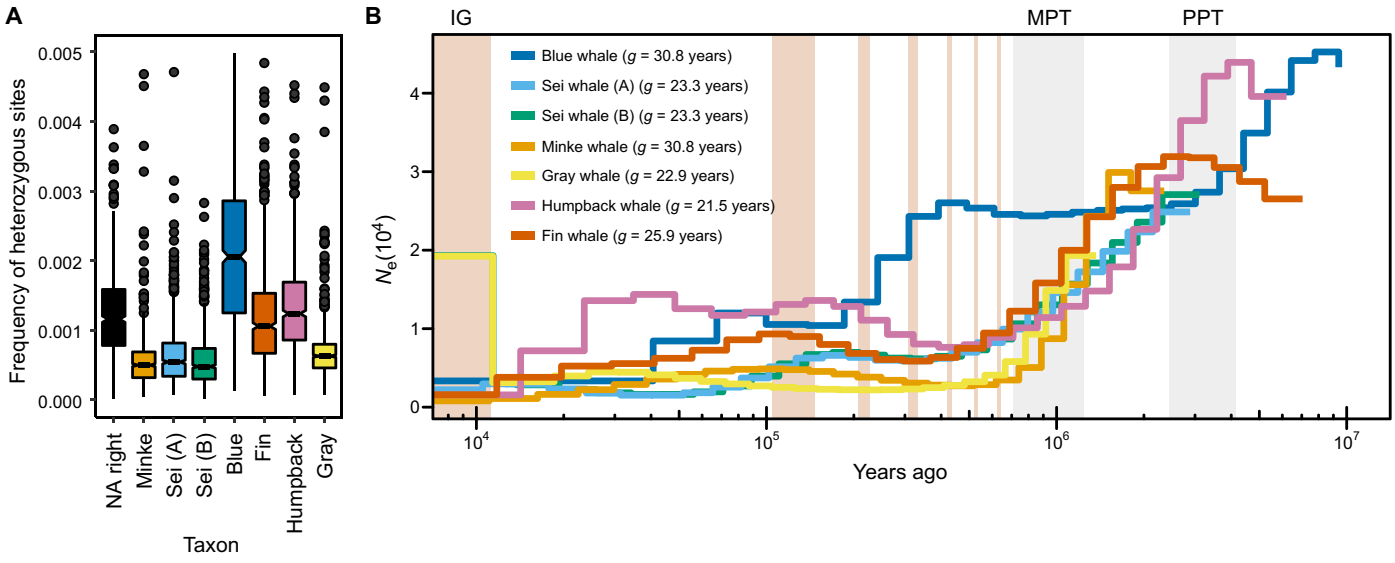

Fig. 4. Demographic history and genome-wide heterozygosity. (**A**) Genome-wide heterozygosity estimated from genomic 100-kbp windows. (**B**) Historical N_e using the PSMC analyses for all baleen whale genomes. The *x* axis shows the time, and the *y* axis shows N_e. Plots were scaled using a mutation rate (μ) of 1.39×10^{-8} substitutions nucleotide^{-1} generation^{-1} and species-specific generation times (*g*). Generation times are noted next to the species names. Light brown shading indicates interglacials (IG) in the Pleistocene and Holocene, and gray shading indicates the MPT and the PPT.

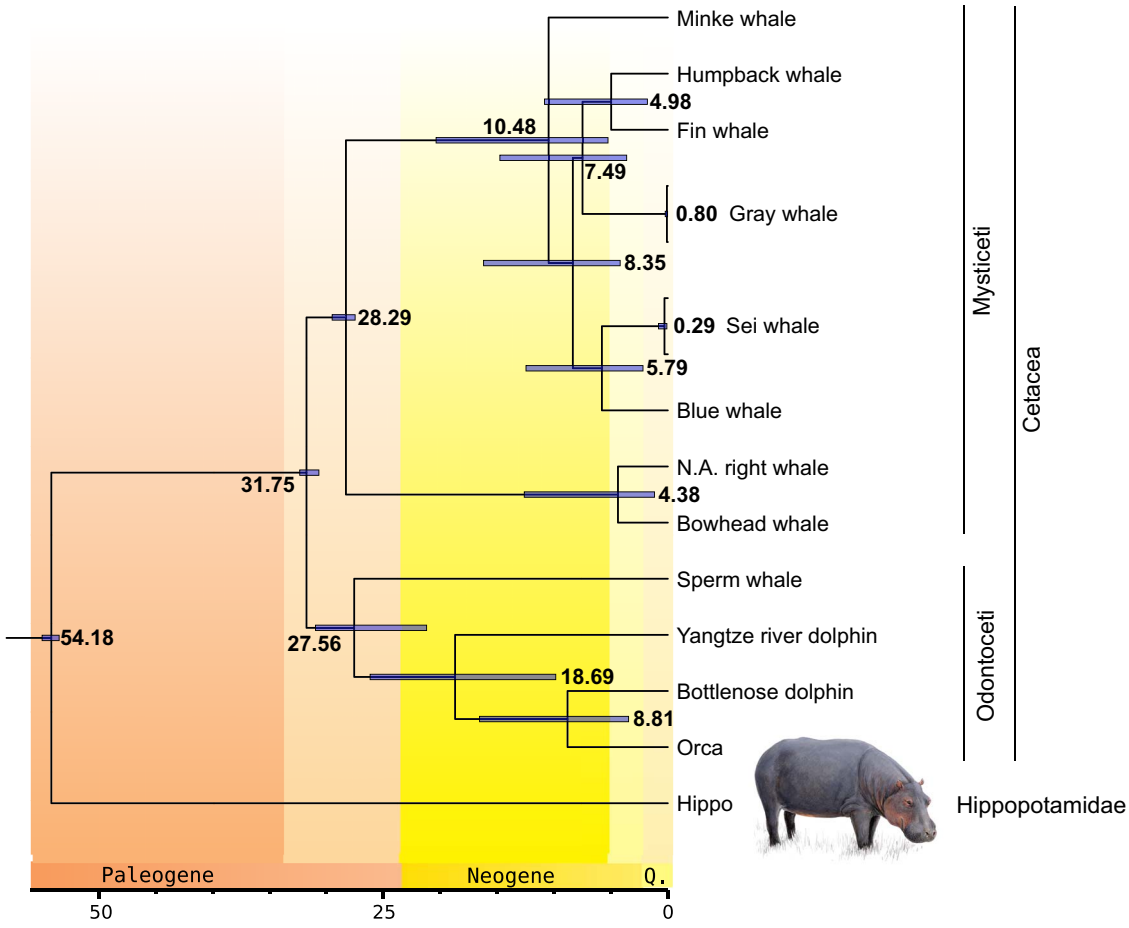

Fig. 5. Divergence time tree of Cetancodonta (*56*) including the newly sequenced baleen whales, estimated from 234,947 amino acid sites (2778 orthologs). Rorquals diverged in the late Miocene, 10.5 to 7.5 Ma ago. Four other cetartiodactyl species were also included but not shown due to space constraints; the dog (*Canis lupus familiaris*) was used as an outgroup. Five calibration points were used for dating (table S8) (*29*, *56*–*60*).

whale, and (iii) fin plus humpback whales. The early rorqual radiation is therefore best understood as a phylogenetic network because different fragments of the rorqual genomes support three different evolutionary histories. This provides the reason why the evolution of rorquals was previously differently reconstructed and poorly supported by molecular analyses of smaller data sets (*5*–*8*). Their evolutionary reconstruction needed to be constrained by morphological data to yield a traditional bifurcating tree among rorquals (*2*).

The apparently unequivocal support for the species tree by the MSC analyses is likely a consequence of a slight imbalance of the evolutionary signal that preferably places the gray whale together with the fin whale and humpback whale. Within the massive amount of genome-scale data, even a minor bias can lead to significantly resolved branches, despite the underlying conflict (*31*). Therefore, inspection of quartet scores in a coalescent species tree and network and CONSENSE analyses are crucial in identifying and depicting conflict in the evolutionary signal.

Rorqual taxonomy

Despite the conflict for the early divergence among rorquals, other divergences are well resolved by genome analyses that find the humpback whale closely related to the fin whale within the genus *Balaenoptera*. This is consistent with previous mitogenomic studies (*5*, *7*, *21*) and makes a separate genus, *Megaptera*, obsolete. If the rules of scientific nomenclature are strictly followed in accord with the phylogenetic relationships, the preferred name of the humpback whale should be *Balaenoptera novaeangliae*.

Because gray whales are morphologically, behaviorally, and ecologically distinct from other balaenopterid whales, placing them in a separate family (Eschrichtiidae) distinct from Balaenopteridae sensu stricto seemed natural (*1*, *32*). This classification has been questioned by some molecular analyses (*5*, *21*), and the current genomic analyses resolve this issue conclusively. Despite their derived morphology, gray whales fall unquestionably within the genus *Balaenoptera*, challenging their status as a separate family or even as a separate genus. Notably, the first described specimen of a gray whale was named *Balaenoptera robusta* (*33*) but later classified as own family and genus by J.E. Gray in 1865 in honor of the zoologist D. F. Eschricht (*32*). Consequently, we suggest that the originally proposed scientific name of the gray whale should be resurrected, with its name included in the Balaenopteridae.

Mechanisms of the rorqual radiation

The radiation of extant rorquals is documented by a rich fossil record with a notable diversity of evolutionary distinct lineages, most of which are now extinct. Speciation is generally assumed to occur when biological or geographic isolation results in reproductive isolation (*34*), and it may be difficult to conceive how whales could diverge. Compared to the terrestrial environment, the marine realm is a three-dimensional continuum, almost devoid of barriers that could aid allopatric speciation for highly mobile organisms such as whales. Mixing of gene pools among rorquals can still occur, and such a process would hinder diversification and consequently speciation (*9*). Even some 8 Ma (or about 400,000 generations ago) after their initial divergence, some baleen whale species can still hybridize, which might also be facilitated by their strikingly uniform karyotypes (*11*).

However, ongoing sympatric speciation in marine mammals by the formation of discrete ecotypes has been suggested for the orca or killer whale (*Orcinus orca*) (*35*). For example, the so-called "transient" and "resident" ecotypes specialized to prey on mammals and fish, respectively (*35*). Similarly, rorquals have evolved different feeding strategies. Whereas most baleen whales feed on pelagic prey such as zooplankton and small fish, the gray whales have evolved to feed on benthic invertebrates by scooping up the seafloor. This opened a new ecological niche to which the gray whale adapted, leading over time to sympatric speciation. The adaptation to the benthic food source also led to notable morphological changes, consequently placing the gray whale into an own family. This differentiation may be triggered by climatic change and other environmental disturbances. These different ecological specializations could have led to a speciation continuum in the past that is similar to the one observed in orcas today.

Genomic analyses find the divergence times of baleen whales to be somewhat younger but within the range of previous estimates (*5*, *8*, *21*). The rorqual radiation coincides with the late Miocene cooling at ~7 Ma ago (*36*). This global cooling affected the marine environment by the onset of the current equator-pole temperature gradient. The beginning of the modern oceanic circulation increased productivity in the temperate and polar oceans (*36*), which may have affected cetacean evolution into different ecotypes.

Network-like evolution in whales

It seems counterintuitive that even whole-genome data do not fully resolve the evolution of whales and other mammals in a bifurcating pattern (*12*). However, speciation being a continuous process with possible hybridization, rather than a strict dichotomous event, has already been recognized by Darwin (*37*) and has recently gained new attention (*38*). In sympatric speciation, genomes can be homogenized by gene flow, and only a few genes need to be under divergent selection to form new species (*38*). Genome analyses sometimes fail to support the idea that speciation by reproductive isolation can fail to yield a fully resolved bifurcating tree, which has been the ultimate goal of evolutionary studies for many years. The analysis of genome sequences rather allows observing and comprehending evolutionary incongruence to translate this into new evolutionary hypotheses that might be better depicted as networks (*39*). Recognizing that "divergence with genetic exchange" is a widespread phenomenon in animals (*9*) makes it necessary to review the biological species concept. Instead of relying on reproductive isolation (*34*), a modern species concept should incorporate selective processes that maintain species divergence even under gene flow (*12*).

Signals for introgressive hybridization

Signals for gene flow confirm sightings and reports of current hybridization in whales (*10*, *40*, *41*). The signal for gene flow between blue and fin whale confirms introgression in these species. Other reports on hybrids between humpback and blue whales (*40*) or between bowhead and right whales (*42*) could not be confirmed by the present genome analyses. The hybridization between these species is likely restricted to few individuals or populations and did not lead to introgression. Further sequencing efforts will give more detailed insights into the extent of introgression of baleen whales and potential ecological implications.

In recent genomic studies of bears, humans, and many other animals, gene flow from introgressive hybridization has been identified as a cause for phylogenetic incongruence (*9*, *12*). Postspeciation gene flow can be analyzed in genomic data with a variety of methods (*43*). The *D* statistic and its derivative are undoubtedly the widest applied methodology (*24*, *25*), but these approaches assume a fully resolved species tree. If the species tree includes polytomies or, based on inappropriate statistical methods, is misidentified (*44*), then the basic assumption of the *D* statistic may be violated and the results can be misleading. Therefore, in case of phylogenetic uncertainties, gene flow

analyses should, in addition, apply methods that do not require a known topology such as PhyloNet that infers introgression signals from a set of gene trees (*26*). However, alternative methods can be computationally intractable for complex phylogenies or a large number of loci.

Demographic history

Genome data from a single individual allow the reconstruction of the effective population size of its species for some 1 to 2 Ma back in time (*28*). These studies have shown that the demographic histories of many mammals have been influenced by climatic oscillations in the Pleistocene [for example, sheep (*45*)]. However, baleen whales maintained relatively stable effective population sizes after the MPT, despite major oscillations in the global climate consequently affecting ocean circulation, upwelling, and marine productivity. The general congruence of population size histories of different baleen whale species indicates that they were similarly affected by these factors. Differences in sequence depth may limit the comparison of absolute N_e between our samples; however, chronology of the curves is not expected to be affected (*46*). Industrial whaling has been too recent to leave a noticeable signal of a declining N_e in the PSMC analyses, especially for long-lived species with long generation times like rorquals. However, compared to other mammals, rorquals, particularly the blue whale, have a comparatively high degree of genome-wide heterozygosity (*27*). The impact of whaling on the genetic diversity of baleen whales may become apparent only after several generations and require population-scale studies for a detailed assessment (*47*).

CONCLUSION

Genome data analyses finally resolved the evolutionary history of baleen whales, even if it is not a bifurcating tree that most had expected. The evolution of rorquals can only be accurately understood by phylogenetic networks because a forced bifurcating tree or a hard polytomy would ignore the accumulated evolutionary history that is recorded in their genomes. It is evident that the central rorqual radiation was not along a progressively ordered process. On the contrary, speciation with gene flow is indicated by the nearly equal probabilities for different evolutionary histories across rorqual genomes. In addition, hybridization between blue and fin whales left genome-wide signals of introgression. The gray whale may constitute a striking example of sympatric speciation related to adaptation to and occupation of a particular niche, bottom feeding, as compared to the pelagic feeding of other rorquals. Our results indicate that sympatric speciation should not be neglected as a mode of speciation in highly connected habitats, such as the marine environment.

MATERIALS AND METHODS

DNA isolation and sequencing

Cell cultures (established by the first author, 1969 to 1974) were grown in Dulbecco's modified Eagle's medium supplemented with 10% fetal bovine serum under standard conditions. DNA of *H. amphibius* was extracted from muscle tissue of a naturally deceased individual, provided by M. Bertelsen (Copenhagen Zoo). DNA was isolated from cells or tissue using a standard phenol-chloroform method. Sequencing libraries were prepared with insert sizes between 300 and 500 bp and sequenced using Illumina HiSeq 2000, 2500, and 4000 technology. The minke whale genome data were obtained from the short read archive (accession no. SRR896642) (*15*). Sequencing library information and mapping statistics are given in table S1. Quality control was performed using FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/), and reads were trimmed. All cell culture work and DNA extractions from tissues were performed according to the ethical guidelines and permission of the respective institutions.

Paired-end reads were mapped to the bowhead whale genome (*B. mysticetus*) (*16*), with BWA mem version 0.7.12-r1039 (*48*), and duplicates were marked with picard (https://github.com/broadinstitute/picard). The bowhead whale was used as reference genome because it avoids a mapping bias that can affect phylogenetic analyses. The minke whale is phylogenetically placed inside baleen whales, and a possible mapping bias against its genome is likely to affect phylogenetic and gene-flow analyses. Scaffolds shorter than 100 kbp were excluded. Repetitive sequences were annotated for the bowhead whale genome by RepeatMasker (*17*). From the mapped reads, single-nucleotide variants (SNVs) and short insertion or deletions (InDels) were called by freebayes v0.9.20-16-g3e35e72 (*49*) with a minimum coverage of four reads and settings: –-monomorphic –min-mapping-quality 20, –C 4, –F 0.3. Consensus sequences were created from VCF-files using custom perl scripts. InDels were removed, and ambiguously called sites were masked as "N."

For sequencing the hippopotamus genome, paired-end and mate-pair libraries were constructed with different insert sizes sequenced on Illumina HiSeq 2000/2500 sequencers (table S4). Because of high levels of duplications, mate-pair libraries were deduplicated. All libraries were trimmed for adaptors and low-quality regions, requiring a minimum read length of 90 bp after trimming. All libraries were assembled into contigs using Minia with $k = 49$ (*19*). Contigs were scaffolded with SSPACE (https://github.com/nsoranzo/sspace_basic) using the mate-pair libraries. Finally, GapCloser (http://soap.genomics.org.cn/) was run with all libraries. Scaffolds shorter than 1 kbp were excluded from the final genome assembly of the hippopotamus. Novel repetitive elements were identified with RepeatModeler (www.repeatmasker.org/RepeatModeler/).

The genome assembly was screened for repetitive sequences using RepeatMasker and the previously created de novo library of identified repeats from RepeatModeler and the RepBase Mammalia library. To account for nonoverlapping detected repeats, we combined and applied the genome masks to the genome sequence. Protein coding genes were predicted ab initio with AUGUSTUS v.3.1 (*20*) using settings –UTR -species human.

Phylogenomic analysis of baleen whales

Consensus sequences of all genomes were aligned per scaffold, and heterozygous sites and repetitive regions were removed. Per-scaffold alignments were split into nonoverlapping GFs of 10, 20, and 100 kbp, respectively. Scaffolds that were shorter than the GF size after removal of ambiguous sites were excluded.

Estimating phylogenetic information in GFs

To analyze the phylogenetic information content of the GFs, we randomly sampled 5000 GFs to count the number of parsimony informative sites and to estimate the genetic distance between the two closest related whales, that is, the bowhead and the North Atlantic right whale. On the basis of real GFs, we simulated GFs between lengths of 1 and 100 kbp to determine which length carries sufficient phylogenetic information to statistically reject alternative topologies (fig. S1). Topology testing was performed using the AU test (*18*).

Species-tree inference and analysis of phylogenetic conflict

JModelTest2 (*50*) identified the suitable nucleotide substitution model by evaluating random 20-kbp GFs. For each GF, phylogenetic trees were

computed with RaxML (*51*) using ML and the GTR + G substitution model that was identified as best fit. Each ML analysis was bootstrapped with 100 replicates. From all 20-kbp GF trees, ASTRAL 4.10.5 (*31*) computed a species tree under the MSC model (exact method) returning quartet scores and posterior probabilities. The species tree was rooted with the bowhead whale and North Atlantic right whale that are outside Balaenopteridae. CONSENSE from the PHYLIP package (*22*) explored conflict among the gene trees by identifying identical splits in a set of given gene trees and summarizing their frequency. Consensus networks of the GF trees were generated using SplitsTree4 (*23*) with different median thresholds. Phylogenetic consensus networks summarize gene tree discordance by drawing alternative edges for each observed split.

Phylogeny of whale mitochondrial genomes

We reconstructed the mitochondrial (mt) genomes from the whale individuals reported herein by mapping the reads to conspecific published mt genomes and generated consensus sequences as described for the nuclear genomes. Mt sequences were aligned to 19 published mt sequences of whales. Accession numbers of mt genomes used as reference for mapping and the phylogenetic analysis are shown in fig. S4. A Bayesian phylogenetic tree was reconstructed using MrBayes version 3.2.2. The analysis was run for 1,200,000 generations with default priors, using the "invgamma" substitution model and an arbitrary burn in of 25% of the samples.

Gene flow analyses

The *D* statistic compares the number of biallelic ABBA and BABA sites in a four-taxon phylogeny and requires a phylogenetic topology following (((H1, H2), H3), O), with H1 to H3 being ingroups and O being the outgroup. For the analyses, the consensus sequences of baleen whales were fragmented into nonoverlapping 100-kbp windows. We applied the *D* statistic to all asymmetric four-taxon phylogenies that can be extracted from the species tree. This resulted in 33 gene flow analyses, such as "(((blue whale, sei whale), fin whale), minke whale)." The direction of gene flow can be estimated in a derivative of the *D* statistic, the D_{FOIL} analysis (*25*), downloaded 15 September 2015 from https://github.com/jbpease/dfoil. The test requires an asymmetric five-taxon tree with a specific topology; therefore, not all combinations of five whale taxa could be analyzed. The D_{FOIL} analyses used the same genomic windows as the *D* statistic analyses.

Our taxon sampling allowed the analysis of the following topologies when considering the estimated species tree as correct because the D_{FOIL} analyses assume a symmetrical five-taxon topology: (i) (((blue, sei), (fin, hump)), NA right); (ii) (((blue, sei), (fin, gray)), minke); (iii) (((blue, sei), (hump, gray)), minke); (iv) (((blue, sei), (hump, gray)), NA right); (v) (((blue, sei), (hump, gray)), bowhead); (vi) (((blue, sei), (fin, gray)), NA right); (vii) (((blue, sei), (fin, gray)), bowhead); (viii) (((blue, sei), (fin, hump)), bowhead); NA right refers to the North Atlantic right whale, whereas the remaining whales are indicated by the first part of their common names.

Maximum likelihood inference for reticulation with PhyloNet

PhyloNet (*26*) is specifically developed to reconstruct reticulated phylogenies from a set of gene trees. We used the ML approach to analyze a set of every 10th GF ML tree, that is, 3419 trees in a coalescent framework that accounts for ILS while allowing different numbers of reticulations (*26*). Subsampling of trees reduced complexity and computational demand. In addition, the bowhead whale, North Atlantic right whale, and sei whale individual "B" were pruned from the input gene trees because their phylogenetic position is unambiguous. The "InferNetwork_ML" method was run with 50 iterations, yielding the five networks with the highest likelihood scores. Analyzing networks with more than one reticulation were too complex and not interpretable from the extended Newick format.

Demographic history

Changes in N_e for the baleen whales were inferred from genome sequences using the PSMC (*28*). We applied PSMC v0.6.5-r67 with input files generated using Samtools mpileup version 1.2 (www.htslib.org) and by applying a minimum mapping and base quality of 30. Using vcfutils, minimum and maximum depth of coverage thresholds were set to 0.5 and 2× the sample's average coverage (table S1). PSMC was run with 25 iterations, an N_0-scaled maximum coalescent time of 20, and a ρ/θ ratio of 5, and the 64 time intervals were parameterized as "4 + 25 × 2 + 4 + 6." PSMC plots were scaled with a mutation rate of $\mu = 4.5 \times 10^{-10}$ mutations bp^{-1} year^{-1} that has been determined for whales (*52*).

Bootstrapping was performed on whole scaffolds. Species-specific predisturbance generation times were used to scale the PSMC plots (*53*). Industrial whaling took place only during the last 200 years, so predisturbance generation times are more accurate for the time frame covered by PSMC. The generation times are shown in Fig. 5.

Genome-wide heterozygosity

To estimate the genome-wide heterozygosity, we randomly sampled 1000 100-kbp nonoverlapping windows for each genome. For these windows, heterozygous SNVs were extracted from the complete set of called variants. Heterozygous sites were excluded if the distance to a called InDel was 10 bp or less or if the sequencing depth at the site was less than 0.5 or 2× the mean sample coverage. This avoids artifacts from assembly errors. For each window, the frequency of heterozygous sites was calculated. In addition, genome-wide heterozygosity and genome-wide sequencing error were inferred using mlRho (*54*). To exclude the potential effects of higher sequencing coverage in the blue whale, the BAM file was downsampled using GATK (genome analysis tool kit) and genome-wide heterozygosity was estimated for ~10× sequencing data.

Cetartiodactyla phylogenomics

Protein sequences for different representative species among Cetartiodactyla were retrieved from ENSEMBL and RefSeq (table S7). For data obtained from RefSeq, Samtools extracted the CDSs from whole-genome sequences using the annotation provided as a General Feature Format (GFF) file.

The annotated CDS for the bowhead whale was used to extract and translate the corresponding genomic regions from baleen whale genomes that were mapped to the bowhead whale Proteinortho version 5.11 screened protein sequences from all genomes listed in table S7. The baleen whale genomes were mapped to the bowhead whale genome and thus their CDSs have the same genomic coordinates. Therefore, the protein sequences of the baleen whales were added after orthology detection based on orthologous proteins identified in the bowhead whale. All proteins for which orthologs were identified in at least nine species were selected, and their sequences were extracted. Protein sequences were aligned individually and trimmed to exclude ambiguously aligned sites. The trimmed alignments were concatenated and used to date the

cetartiodactyl species tree with MCMCTree (*55*) using five calibration points across the tree of Cetartiodactyla (table S8).

SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at http://advances.sciencemag.org/cgi/content/full/4/4/eaap9873/DC1

fig. S1. Possible tree topologies for baleen whales that were evaluated by the AU test.
fig. S2. Phylogenetic content of GFs.
fig. S3. AU test for increasing GF sizes.
fig. S4. MSC-based species trees generated by ASTRAL using 34,192 GFs, with each GF being 20 kbp long.
fig. S5. Phylogenetic tree from mitochondrial genomes for baleen whales.
fig. S6. A majority-rule consensus tree from 34,192 individual GF ML trees (table S6) calculated with the program CONSENSE of the PHYLIP package.
fig. S7. Consensus networks for baleen whales from 34,192 gene trees (10-kbp GF) at different minimum thresholds of gene trees to form an edge.
fig. S8. ML estimates of genome-wide heterozygosity estimated with mlRho.
fig. S9. Blue whale heterozygosity for different sequencing depth.
fig. S10. Demographic histories for each individual whale genome with 100 bootstrap replicates.
table S1. Sequencing and mapping statistics.
table S2. Occurrences of repetitive elements in the bowhead whale genome.
table S3. Number of called substitutions for each whale genome.
table S4. Library and sequencing information for the hippopotamus genome assembly.
table S5. Summary of repetitive elements in the hippopotamus genome.
table S6. A majority-rule consensus analysis of 34,192 individual GF ML trees.
table S7. Common names, scientific names, accession numbers, and source database of additional genomes that were included in the divergence time analyses.
table S8. Calibration points used for the divergence time tree, node age estimates in million years ago, and references.
table S9. Divergence time estimates for Artiodactyla and Cetacea for nodes in the divergence time tree (Fig. 5).
data S1. *D* statistics results.
data S2. D_{FOIL} results.

REFERENCES AND NOTES

1. R. M. Nowak, *Walker's Mammals of the World* (Johns Hopkins Univ. Press, ed. 6, 1999).
2. F. G. Marx, R. E. Fordyce, Baleen boom and bust: A synthesis of mysticete phylogeny, diversity and disparity. *R. Soc. Open Sci.* **2**, 140434 (2015).
3. A. Werth, in *Feeding: Form, Function, and Evolution in Tetrapod Vertebrates*, K. Schwenk, Ed. (Academic Press, 2000), pp. 487–526.
4. G. J. Slater, J. A. Goldbogen, N. D. Pyenson, Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Proc. Biol. Sci.* **284**, 20170546 (2017).
5. A. Hassanin, F. Delsuc, A. Ropiquet, C. Hammer, B. Jansen van Vuuren, C. Matthee, M. Ruiz-Garcia, F. Catzeflis, V. Areskoug, T. T. Nguyen, A. Couloux, Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *C. R. Biol.* **335**, 32–50 (2012).
6. M. Nikaido, H. Hamilton, H. Makino, T. Sasaki, K. Takahashi, M. Goto, N. Kanda, L. A. Pastene, N. Okada, Baleen whale phylogeny and a past extensive radiation event revealed by SINE insertion analysis. *Mol. Biol. Evol.* **23**, 866–873 (2006).
7. T. Sasaki, M. Nikaido, H. Hamilton, M. Goto, H. Kato, N. Kanda, L. Pastene, Y. Cao, R. Fordyce, M. Hasegawa, N. Okada, Mitochondrial phylogenetics and evolution of mysticete whales. *Syst. Biol.* **54**, 77–90 (2005).
8. U. Arnason, A. Gullberg, A. Janke, Mitogenomic analyses provide new insights into cetacean origin and evolution. *Gene* **333**, 27–34 (2004).
9. M. L. Arnold, *Divergence with Genetic Exchange* (Oxford Univ. Press, 2015).
10. R. Spilliaert, G. Vikingsson, U. Arnason, A. Palsdottir, J. Sigurjonsson, A. Arnason, Species hybridization between a female blue whale (*Balaenoptera musculus*) and a male fin whale (*B. physalus*): Molecular and morphological documentation. *J. Hered.* **82**, 269–274 (1991).
11. U. Árnason, I. F. Purdom, K. W. Jones, Conservation and chromosomal localization of DNA satellites in balenopterid whales. *Chromosoma* **66**, 141–159 (1978).
12. V. Kumar, F. Lammers, T. Bidon, M. Pfenninger, L. Kolter, M. A. Nilsson, A. Janke, The evolutionary history of bears is characterized by gene flow across species. *Sci. Rep.* **7**, 46487 (2017).
13. B. M. Hallström, A. Janke, Mammalian evolution may not be strictly bifurcating. *Mol. Biol. Evol.* **27**, 2804–2816 (2010).
14. J. H. Degnan, N. A. Rosenberg, Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends Ecol. Evol.* **24**, 332–340 (2009).
15. H.-S. Yim, Y. S. Cho, X. Guang, S. G. Kang, J.-Y. Jeong, S. S. Cha, H.-M. Oh, J.-H. Lee, E. C. Yang, K. K. Kwon, Y. J. Kim, T. W. Kim, W. Kim, J. H. Jeon, S.-J. Kim, D. H. Choi, S. Jho, H.-M. Kim, J. Ko, H. Kim, Y.-A. Shin, H.-J. Jung, Y. Zheng, Z. Wang, Y. Chen, M. Chen, A. Jiang, E. Li, S. Zhang, H. Hou, T. H. Kim, L. Yu, S. Liu, K. Ahn, J. Cooper, S.-G. Park, C. P. Hong, W. Jin, H.-S. Kim, C. Park, K. Lee, S. Chun, P. A. Morin, S. J. O'Brien, H. Lee, J. Kimura, D. Y. Moon, A. Manica, J. Edwards, B. C. Kim, S. Kim, J. Wang, J. Bhak, H. S. Lee, J.-H. Lee, Minke whale genome and aquatic adaptation in cetaceans. *Nat. Genet.* **46**, 88–92 (2014).
16. M. Keane, J. Semeiks, A. E. Webb, Y. I. Li, V. Quesada, T. Craig, L. B. Madsen, S. van Dam, D. Brawand, P. I. Marques, P. Michalak, L. Kang, J. Bhak, H.-S. Yim, N. V. Grishin, N. H. Nielsen, M. P. Heide-Jørgensen, E. M. Oziolor, C. W. Matson, G. M. Church, G. W. Stuart, J. C. Patton, J. C. George, R. Suydam, K. Larsen, C. López-Otín, M. J. O'Connell, J. W. Bickham, B. Thomsen, J. P. de Magalhães, Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep.* **10**, 112–122 (2015).
17. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-3.0 (2010); www.repeatmasker.org.
18. H. Shimodaira, An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
19. R. Chikhi, G. Rizk, Space-efficient and exact de bruijn graph representation based on a bloom filter, in *Algorithms in Bioinformatics*, B. Raphael, J. Tang, Eds. (Springer, 2012), vol. 7534, pp. 236–248.
20. M. Stanke, O. Schöffmann, B. Morgenstern, S. Waack, Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62 (2006).
21. M. R. McGowen, M. Spaulding, J. Gatesy, Divergence date estimation and a comprehensive molecular tree of extant cetaceans. *Mol. Phylogenet. Evol.* **53**, 891–906 (2009).
22. J. Felsenstein, PHYLIP—Phylogeny inference package (Version 3.2). *Cladistics* **5**, 164–166 (1989).
23. D. H. Huson, D. Bryant, Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.* **23**, 254–267 (2006).
24. E. Y. Durand, N. Patterson, D. Reich, M. Slatkin, Testing for ancient admixture between closely related populations. *Mol. Biol. Evol.* **28**, 2239–2252 (2011).
25. J. B. Pease, M. W. Hahn, Detection and polarization of introgression in a five-taxon phylogeny. *Syst. Biol.* **64**, 651–662 (2015).
26. Y. Yu, J. Dong, K. J. Liu, L. Nakhleh, Maximum likelihood inference of reticulate evolutionary histories. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 16448–16453 (2014).
27. E. Palkopoulou, S. Mallick, P. Skoglund, J. Enk, N. Rohland, H. Li, A. Omrak, S. Vartanyan, H. Poinar, A. Götherström, D. Reich, L. Dalén, Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr. Biol.* **25**, 1395–1400 (2015).
28. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
29. S. Bajpai, P. D. Gingerich, A new Eocene archaeocete (Mammalia, Cetacea) from India and the time of origin of whales. *Proc. Natl. Acad. Sci. U.S.A.* **95**, 15464–15468 (1998).
30. Z. Liu, M. Pagani, D. Zinniker, R. Deconto, M. Huber, H. Brinkhuis, S. R. Shah, R. M. Leckie, A. Pearson, Global cooling during the Eocene-Oligocene climate transition. *Science* **323**, 1187–1190 (2009).
31. E. Sayyari, S. Mirarab, Fast coalescent-based computation of local branch support from quartet frequencies. *Mol. Biol. Evol.* **33**, 1654–1668 (2016).
32. J. E. Gray, Notes on the whalebone-whales; with a synopsis of the species. *Ann. Mag. Nat. Hist.* **3**, 344–350 (1864).
33. W. Lilljeborg, *Öfversigt af de inom Skandinavien (Sverige och Norrige) anträffade Hvalartade Däggdjur (Cetacea)* (1860).
34. E. Mayr, *Animal Species and Evolution* (The Belknap Press of Harvard Univ. Press, 1963).
35. A. D. Foote, N. Vijay, M. C. Ávila-Arcos, R. W. Baird, J. W. Durban, M. Fumagalli, R. A. Gibbs, M. B. Hanson, T. S. Korneliussen, M. D. Martin, K. M. Robertson, V. C. Sousa, F. G. Vieira, T. Vinař, P. Wade, K. C. Worley, L. Excoffier, P. A. Morin, M. T. Gilbert, J. B. W. Wolf, Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nat. Commun.* **7**, 11693 (2016).
36. T. D. Herbert, K. T. Lawrence, A. Tzanova, L. C. Peterson, R. Caballero-Gill, C. S. Kelly, Late Miocene global cooling and the rise of modern ecosystems. *Nat. Geosci.* **9**, 843–847 (2016).
37. C. Darwin, *On the Origin of the Species* (John Murray, London, 1859).
38. J. L. Feder, S. P. Egan, P. Nosil, The genomics of speciation-with-gene-flow. *Trends Genet.* **28**, 342–350 (2012).
39. E. Bapteste, L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J. O. McInerney, D. A. Morrison, L. Nakhleh, M. Steel, L. Stougie, J. Whitfield, Networks: Expanding evolutionary thinking. *Trends Genet.* **29**, 439–441 (2013).
40. P. A. Folkens, R. R. Reeves, B. S. Stewart, P. J. Clapham, J. A. Powell, *Guide to Marine Mammals of the World* (National Audubon Society, 2002).

41. Ú. Árnason, R. Spilliaert, Á. Pálsdóttir, A. Árnason, Molecular identification of hybrids between the two largest whale species, the blue whale (*Balaenoptera musculus*) and the fin whale (*B. physalus*). *Hereditas* **115**, 183–189 (1991).
42. B. P. Kelly, A. Whiteley, D. Tallmon, The Arctic melting pot. *Nature* **468**, 891 (2010).
43. B. A. Payseur, L. H. Rieseberg, A genomic perspective on hybridization and speciation. *Mol. Ecol.* **25**, 2337–2360 (2016).
44. L. Salichos, A. Rokas, Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* **497**, 327–331 (2013).
45. J. Yang, W.-R. Li, F.-H. Lv, S.-G. He, S.-L. Tian, W.-F. Peng, Y.-W. Sun, Y.-X. Zhao, X.-L. Tu, M. Zhang, X.-L. Xie, Y.-T. Wang, J.-Q. Li, Y.-G. Liu, Z.-Q. Shen, F. Wang, G.-J. Liu, H.-F. Lu, J. Kantanen, J.-L. Han, M.-H. Li, M.-J. Liu, Whole-genome sequencing of native sheep provides insights into rapid adaptations to extreme environments. *Mol. Biol. Evol.* **33**, 2576–2592 (2016).
46. K. Nadachowska-Brzyska, R. Burri, L. Smeds, H. Ellegren, PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Mol. Ecol.* **25**, 1058–1072 (2016).
47. R. G. Leduc, F. I. Archer, A. R. Lang, K. K. Martien, B. Hancock-Hanser, J. P. Torres-Florez, R. Hucke-Gaete, H. C. Rosenbaum, K. van Waerebeek, R. L. Brownell Jr., B. L. Taylor, Genetic variation in blue whales in the eastern pacific: Implication for taxonomy and use of common wintering grounds. *Mol. Ecol.* **26**, 740–751 (2017).
48. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
49. E. Garrison, G. Marth, Haplotype-based variant detection from short-read sequencing. arXiv:1207.3907 (2012).
50. D. Darriba, G. L. Taboada, R. Doallo, D. Posada, jModelTest 2: More models, new heuristics and parallel computing. *Nat. Methods* **9**, 772 (2012).
51. A. Stamatakis, RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
52. J. A. Jackson, C. S. Baker, M. Vant, D. J. Steel, L. Medrano-González, S. R. Palumbi, Big and slow: Phylogenetic estimates of molecular evolution in baleen whales (Suborder Mysticeti). *Mol. Biol. Evol.* **26**, 2427–2440 (2009).
53. B. L. Taylor, S. J. Chivers, J. Larese, W. F. Perrin, "Generation length and percent mature estimates for IUCN assessments of cetaceans" (Administrative Report LJ 07-01. National Marine Fisheries Service, Southwest Fisheries Science Centre, 2007).
54. B. Haubold, P. Pfaffelhuber, M. Lynch, MlRho—A program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes. *Mol. Ecol.* **19**, 277–284 (2010).
55. Z. Yang, The BPP program for species tree estimation and species delimitation. *Curr. Zool.* **61**, 854–865 (2015).
56. U. Arnason, A. Gullberg, S. Gretarsdottir, B. M. Ursing, A. Janke, The mitochondrial genome of the sperm whale and a new molecular reference for estimating Eutherian divergence dates. *J. Mol. Evol.* **50**, 569–578 (2000).
57. U. Arnason, A. Gullberg, A. Janke, X. Xu, Pattern and timing of evolutionary divergences among hominoids based on analyses of complete mtDNAs. *J. Mol. Evol.* **43**, 650–661 (1996).
58. E. D. Mitchell, A new cetacean from the late Eocene La Meseta Formation Seymour Island, Antarctic Peninsula. *Can. J. Fish. Aquat. Sci.* **46**, 2219–2235 (1989).
59. R. E. Fordyce, Oligocene origins of skim-feeding right whales: A small archaic balaenid from New Zealand. *J. Vert. Paleontol.* **22**, 54A (2002).
60. J. Gatesy, in *The Timetree of Life*, S. Hedges, S. Kumar, Eds. (Oxford Univ. Press, 2009), pp. 511–515.

Acknowledgments: We are grateful to J. B. Hlidberg (www.fauna.is) for artwork. We thank M. Bertelsen (Zoo Copenhagen) for providing the tissue of the hippopotamus, as well as K. Hildebrandt and L. Olson (University of Alaska, Museum of the North) for giving access to the museum's gray whale tissues (UAM:Mamm:117578 and 99577). We acknowledge Science for Life Laboratory, the National Genomics Infrastructure, and Uppmax for providing assistance in massive parallel sequencing and computational infrastructure. The present study is also a product of the Centre for Translational Biodiversity Genomics (LOEWE-TBG) as part of the "LOEWE—Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" programme of Hesse's Ministry of Higher Education, Research, and the Arts. **Funding:** This study was supported by Hesse's funding program LOEWE and the Leibniz Association, the Royal Physiographic Society, Lund, and Erik Philip-Sörensen's Foundation. **Author contributions:** U.A., F.L., and A.J. conceived the study and scientific objectives. U.A. and A.J. funded genome sequencing. F.L. made the computational analyses with the help of V.K. M.A.N. interpreted the population genetic data. F.L. and A.J. wrote the manuscript with contributions from all authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** Raw sequencing reads for the baleen whales and the hippopotamus have been deposited at the National Center for Biotechnology Information under BioProjects PRJNA389516 and PRJNA389773, respectively. The assembled genome sequence of the hippopotamus is deposited as NKPW00000000. Mitochondrial genomes of newly sequenced baleen whales are deposited in GenBank under accession MF409242-MF409249. All other data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

Submitted 18 September 2017
Accepted 15 February 2018
Published 4 April 2018
10.1126/sciadv.aap9873

Citation: Ú. Árnason, F. Lammers, V. Kumar, M. A. Nilsson, A. Janke, Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Sci. Adv.* **4**, eaap9873 (2018).

Science Advances

Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow

Úlfur Árnason, Fritjof Lammers, Vikas Kumar, Maria A. Nilsson and Axel Janke

Sci Adv **4** (4), eaap9873.
DOI: 10.1126/sciadv.aap9873