bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Speciation in the face of gene flow within the toothed whale superfamily Delphinoidea

Michael V Westbury[1*], Andrea A. Cabrera[1], Alba Rey-Iglesia[1], Binia De Cahsan[1], Stefanie Hartmann[2], Eline D Lorenzen[1*]

1. The GLOBE Institute, University of Copenhagen, Øster Voldgade 5-7, Copenhagen, Denmark
2. Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24-25, Potsdam, Germany

* Corresponding authors: m.westbury@sund.ku.dk, elinelorenzen@sund.ku.dk

Abstract

Understanding speciation is a central aspect in Biology. The formation of new species was once thought to be a simple bifurcation process. However, recent advances in genomic resources now provide the opportunity to investigate the role of post-divergence gene flow in the speciation process. The diversification of lineages in the presence of gene flow appears almost paradoxical. However, with enough time and in the presence of incomplete physical and/or ecological barriers to gene flow, speciation can and does occur. Speciation without complete isolation seems especially likely to occur in highly mobile, wide ranging marine species, such as cetaceans, which face limited geographic barriers. The toothed whale superfamily Delphinoidea represents a good example to further explore speciation in the presence of interspecific gene flow. Delphinoidea consists of three families (Delphinidae, Phocoenidae, and Monodontidae) and within all three families, contemporary interspecific hybrids have been reported. Here, we utilise publicly available genomes from nine species, representing all three families, to investigate signs of post-divergence gene flow across their genomes, and to address the speciation processes that led to the diversity seen today within Delphinoidea. We use a multifaceted approach including: (i) phylogenetics, (ii) the distribution of shared derived alleles, and (iii) demography-based. We find that the divergence and evolution of lineages in Delphinoidea did not follow a simple bifurcating pattern, but were much more complex. Our results indicate multiple, long-lasting ancestral gene flow events both within and among families, which continued for millions of years after initial divergence.

Introduction

The formation of new species involves the divergence of lineages through reproductive isolation. Such isolation can initially occur in allopatry (geographical isolation) or in sympatry (biological/ecological isolation). Over time, these barriers are maintained and strengthened, ultimately leading to the formation of new species (Norris and Hull, 2012). While allopatric speciation requires geographical isolation plus time, sympatric speciation often requires a broader and more complicated set of mechanisms (Turelli et al., 2001). These mechanisms mostly rely on ecologically-mediated natural selection. Parapatric speciation, on

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the other hand, encompasses intermediate scenarios of partial, but incomplete, physical restrictions to gene flow leading to speciation.

Through the analysis of whole-genome datasets, the detection of post-divergence gene flow between distinct species is becoming more commonplace (Árnason et al., 2018; Barlow et al., 2018; Westbury et al., 2020), demonstrating that speciation is much more complex than a simple bifurcating process (Campbell and Poelstra, 2018; Feder et al., 2012). Speciation is not an instantaneous process, but requires tens of thousands to millions of generations to achieve complete reproductive isolation (Butlin and Smadja, 2018; Coyne and Orr, 2004; Liu et al., 2014). The duration it takes to reach this isolation may be especially long in highly mobile marine species, such as cetaceans, due to a relative lack of geographic barriers in the marine realm, and therefore high potential for gene flow (Árnason et al., 2018).

The apparent inability to undergo allopatric speciation in marine species has been termed the marine-speciation paradox (Bierne et al., 2003). However, over the past decade, genomic studies have provided some insights into how speciation can occur within cetaceans (Árnason et al., 2018; Moura et al., 2020). For example, in killer whales (*Orcinus orca*) it has been proposed that initial phases of allopatry may have led to the accumulation of ecological differences between populations, which strengthened population differences even after they came into secondary contact (Foote et al., 2011; Foote and Morin, 2015). However, whether these initial phases of allopatry caused the divergence, or whether speciation occurred purely in sympatry, remains debated (Moura et al., 2015). Yet these two hypotheses are not necessarily mutually exclusive. Instead, differentiation in parapatry, encompassing features of both allopatric and sympatric speciation, may have been key in the evolutionary history of cetaceans.

The toothed whale superfamily Delphinoidea represents an interesting opportunity to further explore speciation in the presence of putative interspecific gene flow. The root of Delphinoidea has been dated to ~19 million years ago (Ma) (95% CI 19.73 - 18.26 Ma) (McGowen et al., 2020) and has given rise to three families: (i) Delphinidae, the most species rich family, which comprises dolphins and ‘black-fish’ (such as killer whales and pilot whales (*Globicephala spp.*)); (ii) Phocoenidae, commonly known as porpoises; and (iii) Monodontidae, which comprises two surviving lineages, belugas (*Delphinapterus leucas*) and narwhals (*Monodon monoceros*).

Delphinoidea is of particular interest, as contemporary interspecific hybrids have been reported within all three families: Delphinidae (Espada et al., 2019; Miyazaki et al., 1992; Silva et al., 2005), Phocoenidae (Willis et al., 2004), and Monodontidae (Skovrind et al., 2019). However, these hybrids represent recent hybridization events that occurred long after species divergence, and their contributions to the parental gene pools is mostly unknown. The presence of more ancient introgressive hybridization events between families, and during the early radiations of these families, has yet to be investigated. With the rapid increase of genomic resources for cetaceans, and in particular for species within Delphinoidea, we are

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

presented with the ideal opportunity to investigate post-divergence gene flow between lineages, furthering our understanding of speciation processes in cetaceans.

Here, we utilise publicly available whole-genome data from nine species of Delphinoidea, representing all three families, to investigate signs of post-divergence gene flow across their genomes. Our analyses included five Delphinidae (killer whale, Pacific white-sided dolphin (*Lagenorhynchus obliquidens*), long-finned pilot whale (*Globicephala melas*), bottlenose dolphin (*Tursiops truncatus*), Indo-Pacific bottlenose dolphin (*T. aduncus*)); two Phocoenidae (harbour porpoise (*Phocoena phocoena*), finless porpoise (*Neophocaena phocaenoides*)); and two Monodontidae (beluga, narwhal). Moreover, we compare their species-specific genetic diversity and demographic histories, and explore how species abundances may have played a role in interspecific hybridisation over the last two million years.

Results and discussion

Detecting gene flow

To assess the evolutionary relationships across the genomes of the nine Delphinoidea species investigated, we computed non-overlapping sliding-window maximum-likelihood phylogenies of four different window sizes in RAxML (Stamatakis, 2014). These analyses resulted in 43,207 trees (50 kilobase (kb) windows), 21,387 trees (100 kb windows), 3,705 trees (500 kb windows), and 1,541 trees (1 megabase (Mb) windows) (Fig. 1, Supplementary Fig. S1). Regardless of window size, we retrieve consensus support for the species tree previously reported using target sequence capture (McGowen et al., 2020). However, when considering the smallest window size (50 kb), we find a considerable proportion of trees (up to 76%) with an alternative topology to the known species tree (Fig. 1A). These alternative topologies could be due to incomplete lineage sorting (ILS) or interspecific gene flow (Leaché et al., 2014). Moreover, the higher prevalence of this pattern in the 50 kb windows (21% of windows show an alternative topology in the 1 Mb dataset (Fig. 1B)), may indicate that inconsistencies in topology are caused by ancient, rather than recent, events.

To further explore potential gene flow while taking ILS into account, we applied D-statistics. D-statistics uses a four-taxon approach [[[H1, H2], H3], Outgroup] to uncover the differential distribution of shared derived alleles, which may represent gene flow between either H1/H3 or H2/H3. Here we used baiji (*Lipotes vexillifer*) as the outgroup, and alternated the ingroup positions based on the consensus topology. We find that 85 out of 86 tests show signs of gene flow within and between families (Supplementary table S1), suggesting the evolutionary history of Delphinoidea is very complex.

Due to the inability of the four-taxon D-statistics approach to detect the direction of gene flow, as well as whether gene flow events may have occurred between ancestral lineages, we used D-foil. D-foil enables further characterization of the D-statistics results, which may be particularly relevant, given the complex array of gene flow putatively present within Delphinoidea. D-foil uses a five-taxon approach [[H1, H2] [H3, H4], Outgroup] and a

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

system of four independent D-statistics in a sliding-window fashion to uncover (i) putative gene flow events, (ii) donor and recipient lineages, and (iii) whether gene flow events occurred between a distantly related lineage and the ancestor of two sister lineages, which is indicative of ancestral lineage gene flow. However, due to the input topology requirements of D-foil, we were only able to investigate gene flow between families, and not within families, using this analysis. Hence, we tested for gene flow between Delphinidae/Phocoenidae, Delphinidae/Monodontidae, and Monodontidae/Phocoenidae.

The D-foil results underscore the complex pattern of post-divergence gene flow between families indicated by the D-statistics. We find support for interfamilial gene flow events between all nine species investigated, to varying extents (Supplementary table S2). This could reflect multiple episodes of gene flow between all investigated species. Alternatively, the pattern could reflect ancient gene flow events between the ancestors of H1-H2 and H3-H4 (in the topology [[H1, H2] [H3, H4], Outgroup]), with differential inheritance of the admixed loci in subsequent lineages. Such ancestral gene flow events have previously been shown to lead to false positives between species pairs using D-statistics (Moodley et al., 2020). A further putative problem with these results can be seen when implementing D-foil on the topology [[Delphinidae, Delphinidae], [Monodontidae, Phocoenidae], Outgroup]. We find the majority of windows support a closer relationship between Delphinidae (ancestors of H1 and H2) and Monodontidae (H3), as opposed to the species tree. If this result is correct, it suggests the input topology was incorrect, implying that Delphinidae and Monodontidae are sister lineages, as opposed to Phocoenidae and Monodontidae. However, this contrasts with the family topology of [Delphinidae, [Phocoenidae, Monodontidae]] retrieved in our phylogenetic analyses (Fig. 1) and reported by others (McGowen et al., 2020; Steeman et al., 2009). Instead, we suggest our result reflects the limited ability of D-foil to infer gene flow between these highly divergent lineages.

False positives and potential biases in D-statistics and D-foil can arise due to a number of factors including (i) ancestral population structure, (ii) introgression from unsampled and/or extinct ghost lineages, (iii) differences in relative population size of lineages or in the timing of gene flow events, (iv) different evolutionary rates or sequencing errors between H1 and H2, and (v) gene flow between ancestral lineages (Moodley et al., 2020; Slatkin and Pollack, 2008; Zheng and Janke, 2018). These issues are important to consider when interpreting our results, as the deep divergences of lineages suggest there were probably a number of ancestral gene flow events, as well as gene flow events between now-extinct lineages, that may bias results.

Cessation of gene flow

To further elucidate the complexity of interspecific gene flow within Delphinoidea, we implemented F1 hybrid PSMC (hPSMC) (Cahill et al., 2016). This method creates a pseudo-diploid sequence by merging pseudo-haploid sequences from two different genomes, which in our case represents two different species. The variation in the interspecific pseudo-F1 hybrid genome cannot coalesce more recently than the emergence of reproductive

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

isolation between the two parental species, and the method can therefore be used to infer when gene flow between species ceased.

When considering the uppermost limit of when gene flow ended (equating to the most ancient date) and the lower confidence interval of each divergence date (equating to the most recent date), the majority of comparisons (29/36) show that gene flow continued for >50% of the post-divergence branch length (Fig. 2, Supplementary results). This finding suggests that reaching complete reproductive isolation in Delphinoidea was a slow process. Long-term, continuous gene flow may reflect the ability of these cetacean species to travel long distances, and the lack of significant geographical barriers in the marine environment.

Despite our finding of long-term, continued gene flow in the majority of comparisons, our results suggest gene flow ceased more rapidly within Delphinidae, relative to within Phocoenidae and Monodontidae (Fig. 2). Only three out of ten pairwise comparisons (killer whale vs. Indo-Pacific white-sided dolphin, killer whale vs long-finned pilot whale, and bottlenose dolphin vs Indo-Pacific bottlenose dolphin), showed continued gene flow for >50% of the branch length post divergence. The remaining seven comparisons showed continued gene flow along 48% - 24% of the post-divergence branch length. This finding may reflect the inability of hPSMC to detect low levels of migration until the present day, leading to large estimated intervals around the time point at which gene flow stopped.

Simulations have shown that in the presence of only 1/10,000 migrants per generation, hPSMC suggests continued gene flow. However, this does not happen with a lower rate of ~1/100,000 migrants per generation. Rather, in the latter case, the exponential increase in Ne of the pseudo-hybrid genome, which is used to infer the date at which gene flow ceased between the parental individuals, becomes a more gradual transition, leading to a larger estimated time interval (Cahill et al., 2016). Within Delphinidae, we observe a corresponding, less pronounced increase in Ne in the pseudo-hybrids, suggestive of continued, but very low migration rates (Supplementary results). This finding suggests gene flow within Delphinidae may have continued for longer than shown by hPSMC, which may not be sensitive enough to detect the low rates of recent gene flow. Furthermore, persistent gene flow is supported by confirmed fertile contemporary hybrids between some of our study species; for example, bottlenose dolphins can produce fertile offspring with both Indo-Pacific bottlenose dolphins (Gridley et al., 2018) and Pacific white-sided dolphins (Crossman et al., 2016; Miyazaki et al., 1992). Either way, our hPSMC results within and between all three families show a consistent pattern of long periods of interspecific migration in Delphinoidea, some lasting up to more than ten million years post divergence.

Interspecific hybridisation

Interspecific hybridization may occur at a higher rate during periods of low abundance, when a given species encounters only a limited number of conspecifics (Crossman et al., 2016; Edwards et al., 2011; Westbury et al., 2019), and individuals may mate with a closely related species instead of investing energy in finding a rarer conspecific mate. To explore the relationship between susceptibility to interspecific hybridisation and

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

population size, we calculated the level of genome-wide genetic diversity for each species, as a proxy for their population size (Fig. 3A). Narwhal, killer whale, beluga and long-finned pilot whale have the lowest diversity levels, and should therefore be more susceptible to interspecific hybridization events. A beluga/narwhal hybrid has been reported (Skovrind et al., 2019), as has hybridisation between long-finned and short-finned pilot whales (Miralles et al., 2016). However, hybrids between species with high genetic diversity, including harbour porpoise (Willis et al., 2004), Indo-Pacific bottlenose dolphin (Baird et al., 2012), and bottlenose dolphin (Espada et al., 2019; Herzingl and Johnsonz, 1997) have also been reported, suggesting genetic diversity alone is not a good proxy for susceptibility to hybridisation.

To investigate whether interspecific gene flow took place during past periods of low population size, we estimated changes in intraspecific genetic diversity through time (Fig. 3B-D). The modeled demographic trajectories span the past two million years using a Pairwise Sequentially Markovian Coalescent model (PSMC). We could therefore assess the relationship for the three species pairs where the interval for the cessation of gene flow was contained within this period: harbour/finless porpoise (Phocoenidae), beluga/narwhal (Monodontidae), and bottlenose/Indo-pacific bottlenose dolphin (Delphinidae) (Fig. 2).

In the harbour porpoise, we observe an increase in effective population size (Ne) beginning ~1 Ma, the rate of which increases further ~500 thousands of years ago (kya) (Fig. 3C). The timing of expansion overlaps the period during which gene flow with the finless porpoise ceased (~1.1 - 0.5 Ma, Fig. 2), suggesting gene flow between the two species occurred when population size in the harbour porpoise was lower. We observe a similar pattern in belugas; an increase in Ne ~1 Ma, relatively soon after the proposed cessation of gene flow with narwhals ~1.8 - 1.2 Ma (Fig. 3D). An increase in Ne may coincide with an increase in relative abundance, which would increase the number of potential conspecific mates, and in turn reduce the level of interspecific gene flow. Although we are unable to test the direction and levels of gene flow between these species pairs, we expect a relative reduction of gene flow into the more abundant species. A relative reduction of such events would in turn lessen genomic signs of interspecific gene flow, despite its occurrence.

We observe a different pattern in the bottlenose/Indo-pacific bottlenose dolphins. In the previous examples, we find relatively low population size when gene flow was ongoing, and only in one of the two hybridizing species. In the dolphins, we find relatively high population size during the period of gene flow in both species; Ne declines ~1 - 0.5 Ma, coinciding with the putative end of gene flow ~1.2 - 0.4 Ma. The decline in Ne could either reflect a decline in abundance, or a loss of connectivity between the two species. In the latter, we expect levels of intraspecific diversity (and thereby inferred Ne) to decline with the cessation of gene flow, even if absolute abundances did not change. This is indeed suggested by our data, which shows both species undergoing the decline simultaneously, indicative of a common cause.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Seven of the nine Delphinoidea genomes investigated show a similar pattern of a rapid decline in Ne starting ~150 - 100 kya (Fig. 3B-D; the exceptions are narwhal and Pacific white-sided dolphin). This concurrent decline could represent actual population declines across species, or alternatively, simultaneous reductions in connectivity among populations within each species. Based on similar PSMC analyses, a decline in Ne at this time has also been reported in four baleen whale species (Árnason et al., 2018). Although this could reflect demographic factors, such as the loss of population connectivity, the unique life histories, distributions, and ecology of these cetacean species suggests that decreased population connectivity is unlikely to have occurred simultaneously across all studied species. Rather, the species-wide pattern may reflect climate-driven environmental change.

The period 150 - 100 kya overlaps with the onset of the last interglacial, when sea levels increased to levels as high, if not higher, than at present (Polyak et al., 2018), and which may have had a marine-wide effect on population sizes. A similar marine-wide effect has been observed among baleen whales and their prey species in the Southern and North Atlantic Oceans during the Pleistocene-Holocene climate transition (12-7 kya) (Cabrera et al., 2018). These results lend support to the ability of marine-wide environmental shifts to drive changes in population sizes across multiple species.

We suggest species-wide declines may have facilitated the resurgence of hybridization between the nine Delphinoidea species analysed here. If hybridisation did increase, species may already have been sufficiently differentiated that offspring fertility was reduced. Even if offspring were fertile, the high level of differentiation between species may have meant hybrids were unable to occupy either parental niche (Skovrind et al., 2019) and were therefore strongly selected against. A lack of significant contribution from hybrids to the parental gene pools may be why we observe contemporary hybrids, despite lacking evidence of this in the hPSMC analysis.

Conclusions

Allopatric speciation is generally considered the most common mode of speciation, as the absence of gene flow due to geographical isolation can most easily explain the evolution of ecological, behavioral, morphological, or genetic differences between populations (Norris and Hull, 2012). However, our findings suggest that within Delphinoidea, speciation in the presence of gene flow was commonplace, consistent with sympatric/parapatric speciation.

Long periods of extended post-divergence gene flow may also explain the presence of contemporaneous hybrids between several species. In parapatric speciation, genetic isolation is achieved relatively early due to geographical and biological isolation, but species develop complete reproductive isolation relatively slowly, allowing hybridization to continue for an extended period of time (Norris and Hull, 2012). The prevalence of this mode of speciation in cetaceans, as suggested by our study and previous genomic analyses (Árnason et al., 2018; Moura et al., 2020), may reflect the low energetic costs of dispersing across large distances in the marine realm (Fish et al., 2008; Williams, 1999) and the relative absence of geographic

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

barriers preventing such dispersal events (Palumbi, 1994). Both factors are believed to be important in facilitating long-distance (including inter-hemispheric and inter-oceanic) movements in many cetacean species (Stone et al., 1990).

Methods

Data collection

We downloaded the assembled genomes and raw sequencing reads from nine toothed whales from the superfamily Delphinoidea. The data included five Delphinidae: Indo-Pacific white-sided dolphin (NCBI Biosample: SAMN09386610), Indo-Pacific bottlenose dolphin (NCBI Biosample: SAMN06289676), bottlenose dolphin (NCBI Biosample: SAMN09426418), killer whale (NCBI Biosample: SAMN01180276), and long-finned pilot whale (NCBI Biosample: SAMN11083132); two Phocoenidae: harbour porpoise (available from Autenrieth et al., 2018), finless porpoise (NCBI Biosample: SAMN02192673); and two Monodontidae: beluga (NCBI Biosample: SAMN06216270), narwhal (NCBI Biosample: SAMN10519625). To avoid biases that may occur when mapping to an ingroup reference (Westbury et al., 2019), we used the assembled baiji genome (Genbank accession code: GCF_000442215.1) as mapping reference in the gene flow analyses. Delphinoidea and the baiji diverged ~24.6 Ma (95% CI 25.2 - 23.8 Ma) (McGowen et al., 2020).

Initial data filtering

To determine which scaffolds were most likely autosomal in origin, we identified putative sex chromosome scaffolds for each genome, and omitted them from further analysis. We found putative sex chromosome scaffolds in all ten genomes by aligning the assemblies to the Cow X (Genbank accession: CM008168.2) and Human Y (Genbank accession: NC_000024.10) chromosomes. Alignments were performed using satsuma synteny v2.1 (Grabherr et al., 2010) with default parameters. We also removed scaffolds smaller than 100 kb from all downstream analyses.

Mapping

We trimmed adapter sequences from all raw reads using skewer v0.2.2 (Jiang et al., 2014). We mapped the trimmed reads to the baiji for downstream gene flow analyses, and to the species-specific reference genome for downstream demographic history and genetic diversity analyses using BWA v0.7.15 (Li and Durbin, 2009) and the mem algorithm. We parsed the output and removed duplicates and reads with a mapping quality lower than 30 with SAMtools v1.6 (Li et al., 2009). Mapping statistics can be found in supplementary tables S3 and S4.

Sliding-window phylogeny

For the sliding-window phylogenetic analysis, we created fasta files for all individuals mapped to the baiji genome using a consensus base call (-dofasta 2) approach in ANGSD v0.921 (Korneliussen et al., 2014), and specifying the following filters: minimum read depth of 5 (-mininddepth 5), minimum mapping quality of 30 (-minmapq 30), minimum base quality (-minq 30), only consider reads that map to one location uniquely (-uniqueonly 1),

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

and only include reads where both mates map (-only_proper_pairs 1). All resultant fasta files, together with the assembled baiji genome, were aligned, and sites where any individual had more than 50% missing data were filtered before performing maximum likelihood phylogenetic analyses in a non-overlapping sliding-window approach using RAxML v8.2.10 (Stamatakis, 2014). We performed this analysis four times independently, specifying a different window size each time (50 kb, 100 kb, 500 kb, and 1 Mb). We used RAxML with default parameters, specifying baiji as the outgroup, and a GTR+G substitution model. We computed the genome-wide majority rule consensus tree for each window size in PHYLIP (Felsenstein, 2005), with branch support represented by the proportion of trees displaying the same topology. We simultaneously visualised all trees of the same sized window using DensiTree (Bouckaert, 2010).

D-statistics

To test for signs of gene flow in the face of incomplete lineage sorting (ILS), we ran D-statistics using all individuals mapped to the baiji genome in ANGSD, using a consensus base call approach (-doabbababa 2), specifying the baiji sequence as the ancestral outgroup sequence, and the same filtering as for the fasta file construction with the addition of setting the block size as 1Mb (-blocksize). Significance of the results was evaluated using a block jackknife approach with the Rscript provided in the ANGSD package. $|Z| > 3$ was deemed significant.

D-foil

As D-statistics only tests for the presence and not the direction of gene flow, we ran D-foil (Pease and Hahn, 2015), an extended version of the D-statistics, which is a five-taxon test for gene flow, making use of all four combinations of the potential D-statistics topologies. For this analysis, we used the same fasta files constructed above, which we converted into an mvf file using MVFtools (Pease and Rosenzweig, 2018). We specified the 5-taxon [[H1, H2], [H3, H4], baiji], for all possible combinations, following the species tree (Fig. 1) and a 100 kb window size. All scaffolds were trimmed to the nearest 100 kb to avoid the inclusion of windows shorter than 100 kb.

Mutation rate estimation

For use in the downstream demographic analyses, we computed the mutation rate per generation for each species. To do this, we estimated the pairwise distances between all ingroup species mapped to the baiji, using a consensus base call in ANGSD (-doIBS 2), and applying the same filters as above, with the addition of only considering sites in which all individuals were covered (-minInd). The pairwise distances used in this calculation were those from the closest lineage to the species of interest (Supplementary tables S5 and S6). The mutation rates per generation were calculated using the resultant pairwise distance as follows: mutation rate = pairwise distance x generation time / 2 x divergence time. Divergence times were taken from the full dataset 10-partition AR (mean) values from McGowen et al. (McGowen et al., 2020) (Supplementary table S6). Generation times were taken from previously published data (Supplementary table S7).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Cessation of gene flow

To estimate when gene flow may have ceased between each species pair, we used the F1-hybrid PSMC (hPSMC) approach (Cahill et al., 2016). As input we used the haploid consensus sequences created for the phylogenetic analyses. We merged the haploid sequences from each possible species pair into pseudo-diploid sequences using the scripts available in the hPSMC toolsuite. We independently ran each resultant species pair pseudo-diploid sequences through PSMC, specifying atomic intervals 4+25*2+4+6. We plotted the results using the average (i) mutation rate per generation and (ii) generation time for each species pair being tested. From the output of this analysis, we visually estimated the pre-divergence Ne of each hPSMC plot (i.e. Ne prior to the point of asymptotic increase in Ne) to be used as input for downstream simulations. Based on these empirical results, we ran simulations in ms (Hudson, 2002) using the estimated pre-divergence Ne, and various predefined divergence times to find the interval in which gene flow may have ceased between a given species pair. The time intervals and pre-divergence Ne for each species pair used for the simulations can be seen in supplementary table S8. The ms commands were produced using the scripts available in the hPSMC toolsuite. We plotted the simulated and empirical hPSMC results to find the simulations with an asymptotic increase in Ne closest to, but not overlapping with, the empirical data. The predefined divergence times of the simulations showing this pattern within 1.5x and 10x of the pre-divergence Ne were taken as the time interval in which gene flow ceased.

Heterozygosity

As a proxy for species-level genetic diversity, we estimated autosome-wide heterozygosity for each of the nine Delphinoidea species. We estimated autosomal heterozygosity using allele frequencies (-doSaf 1) in ANGSD (Korneliussen et al., 2014), taking genotype likelihoods into account (-GL 2) and specifying the same filters as for the fasta file construction with the addition of adjusting quality scores around indels (-baq 1), and the subsample filter (-downSample), which was uniquely set for each individual to result in a 20x genome-wide coverage, to ensure comparability between genomes of differing coverage. Heterozygosity was computed from the output of this using realSFS from the ANGSD toolsuite and specifying 20 Mb windows of covered sites (-nSites).

Demographic reconstruction

To determine the demographic histories of all nine species over a two million year time scale, we ran a Pairwise Sequentially Markovian Coalescent model (PSMC) (Li and Durbin, 2011) on each diploid genome independently. We called diploid genome sequences using SAMtools and BCFtools v1.6 (Narasimhan et al., 2016), specifying a minimum quality score of 20 and minimum coverage of 10. We ran PSMC specifying atomic intervals 4+25*2+4+6 and performed 100 bootstrap replicates to investigate support for the resultant demographic trajectories. PSMC outputs were plotted using species-specific mutation rates and generation times (Supplementary table S7).

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure legends:

Figure 1: Sliding-Window Maximum likelihood trees of nine Delphinoidea species and the baiji. Simultaneously plotted trees constructed using non-overlapping sliding windows of (A) 50 kb in length and (B**)** 1 Mb in length. Black lines show the consensus tree. Grey lines show individual trees. Numbers on branches show the proportion of windows supporting the node. Branches without numbers show 100% support. Baiji, killer whale, white-sided dolphin, pilot whale, harbour porpoise, finless whale, beluga, and narwhal silhouettes: Chris huh, license CC-BY-SA-3.0 (https://creativecommons.org/licenses/by-sa/3.0/). Bottlenose dolphin silhouette: license Public Domain Dedication 1.0.

Figure 2: Estimated divergence times and time interval during which gene flow ceased between species (A) within families and (B) between families. Estimated time intervals of when gene flow ceased between species pairs are based on hPSMC results and simulated data. Divergence time estimates are taken from the full dataset 10-partition AR results of McGowen et al 2020.

Figure 3: Autosome-wide heterozygosity and demographic histories over the last two million years. (A) Autosome-wide levels of heterozygosity calculated in 20 Mb windows of consecutive bases. (B-D) Demographic history of all studied species within (B) Delphinidae, (C) Phocoenidae, and (D) Monodontidae, estimated using PSMC. Thick coloured lines show the autosome-wide demographic history. Faded lines show bootstrap support values.

Acknowledgements

The work was supported by the Carlsberg Foundation Distinguished Associate Professor Fellowship, grant no CF16-0202, the Villum Fonden Young Investigator Programme, grant no. 13151, and the Independent Research Fund Denmark | Natural Sciences, Forskningsprojekt 1, grant no. 8021-00218B to EDL. AAC was funded by the Rubicon-NWO grant (project 019.183EN.005). We would like to thank all those contributing to the ever-increasing abundance of publicly available genomic resources. Without the availability of such data, our study would not have been possible.

Author contributions

Conceptualization, MVW; Formal analysis, MVW, AAC, AR-I, BDC, SH; Writing – Original Draft MVW; Writing – Review & Editing, All authors; Supervision, MVW, EDL; Funding Acquisition, EDL.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

References:

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

Autenrieth M, Hartmann S, Lah L, Roos A, Dennis AB, Tiedemann R. 2018. High-quality whole-genome sequence of an abundant Holarctic odontocete, the harbour porpoise (*Phocoena phocoena*). *Mol Ecol Resour* **18**:1469–1481.

Baird RW, Gorgone AM, McSweeney DJ, Ligon AD, Deakos MH, Webster DL, Schorr GS, Martien KK, Salden DR, Mahaffy SD. 2012. Population structure of island associated dolphins: Evidence from mitochondrial and microsatellite markers for common bottlenose dolphins (*Tursiops truncatus*) in the main Hawaiian Islands. *Mar Mamm Sci.*

Barlow A, Cahill JA, Hartmann S, Theunert C, Xenikoudakis G, Fortes GG, Paijmans JLA, Rabeder G, Frischauf C, Grandal-d'Anglade A, García-Vázquez A, Murtskhvaladze M, Saarma U, Anijalg P, Skrbinšek T, Bertorelle G, Gasparian B, Bar-Oz G, Pinhasi R, Slatkin M, Dalén L, Shapiro B, Hofreiter M. 2018. Partial genomic survival of cave bears in living brown bears. *Nat Ecol Evol* **2**:1563–1570.

Bierne N, Bonhomme F, David P. 2003. Habitat preference and the marine-speciation paradox. *Proc Biol Sci* **270**:1399–1406.

Bouckaert RR. 2010. DensiTree: making sense of sets of phylogenetic trees. *Bioinformatics* **26**:1372–1373.

Butlin RK, Smadja CM. 2018. Coupling, Reinforcement, and Speciation. *Am Nat* **191**:155–172.

Cabrera AA, Schall E, Bérubé M, Bachmann L, Berrow S, Best PB, Clapham PJ, Cunha HA, Rosa LD, Dias C, Findlay KP, Haug T, Heide-Jørgensen MP, Kovacs KM, Landry S, Larsen F, Lopes XM, Lydersen C, Mattila DK, Oosting T, Pace RM, Papetti C, Paspati A, Pastene LA, Prieto R, Ramp C, Robbins J, Ryan C, Sears R, Secchi ER, Silva MA, Víkingsson G, Wiig Ø, Øien N, Palsbøll PJ. 2018. Strong and lasting impacts of past global warming on baleen whale and prey abundance. *bioRxiv*.

Cahill JA, Soares AER, Green RE, Shapiro B. 2016. Inferring species divergence times using pairwise sequential Markovian coalescent modelling and low-coverage genomic data. *Philos Trans R Soc Lond B Biol Sci* **371**. doi:10.1098/rstb.2015.0138

Campbell CR, Poelstra JW. 2018. What is Speciation Genomics? The roles of ecology, gene flow, and genomic architecture in the formation of species. *Biol J Linn Soc Lond* **124**:561–583.

Coyne JA, Orr HA. 2004. Speciation. Sinauer Associates Sunderland, MA.

Crossman CA, Taylor EB, Barrett-Lennard LG. 2016. Hybridization in the Cetacea: widespread occurrence and associated morphological, behavioral, and ecological factors. *Ecol Evol* **6**:1293–1303.

Edwards CJ, Suchard MA, Lemey P, Welch JJ, Barnes I, Fulton TL, Barnett R, O'Connell TC, Coxon P, Monaghan N, Valdiosera CE, Lorenzen ED, Willerslev E, Baryshnikov GF, Rambaut A, Thomas MG, Bradley DG, Shapiro B. 2011. Ancient hybridization and an Irish origin for the modern polar bear matriline. *Curr Biol* **21**:1251–1258.

Espada R, Olaya-Ponzone L, Haasova L, Martín E, García-Gómez JC. 2019. Hybridization in the wild between *Tursiops truncatus* (Montagu 1821) and *Delphinus delphis* (Linnaeus 1758). *PLoS One* **14**:e0215020.

Feder JL, Egan SP, Nosil P. 2012. The genomics of speciation-with-gene-flow. *Trends Genet* **28**:342–350.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6.

Fish FE, Howle LE, Murray MM. 2008. Hydrodynamic flow control in marine mammals.

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Integr Comp Biol **48**:788–800.
Foote AD, Morin PA. 2015. Sympatric speciation in killer whales? *Heredity* **114**:537–538.
Foote AD, Morin PA, Durban JW, Willerslev E. 2011. Out of the Pacific and back again: insights into the matrilineal history of Pacific killer whale ecotypes. *PLoS*.
Grabherr MG, Russell P, Meyer M, Mauceli E, Alföldi J, Di Palma F, Lindblad-Toh K. 2010. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**:1145–1151.
Gridley T, Elwen SH, Harris G, Moore DM, Hoelzel AR, Lampen F. 2018. Hybridization in bottlenose dolphins—A case study of *Tursiops aduncus* × *T. truncatus* hybrids and successful backcross hybridization events. *PLoS One* **13**:e0201722.
Herzingl DL, Johnsonz CM. 1997. Interspecific interactions between Atlantic spotted dolphins (*Stenella frontalis*) and bottlenose dolphins (*Tursiops truncatus*) in the Bahamas 1985-1995. *Aquat Mamm*.
Hudson RR. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**:337–338.
Jiang H, Lei R, Ding S-W, Zhu S. 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**:182.
Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**:356.
Leaché AD, Harris RB, Rannala B, Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. *Syst Biol* **63**:17–30.
Li H, Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* **475**:493–496.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**:1754–1760.
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**:2078–2079.
Liu S, Lorenzen ED, Fumagalli M, Li B, Harris K, Xiong Z, Zhou L, Korneliussen TS, Somel M, Babbitt C, Wray G, Li J, He W, Wang Z, Fu W, Xiang X, Morgan CC, Doherty A, O'Connell MJ, McInerney JO, Born EW, Dalén L, Dietz R, Orlando L, Sonne C, Zhang G, Nielsen R, Willerslev E, Wang J. 2014. Population genomics reveal recent speciation and rapid evolutionary adaptation in polar bears. *Cell* **157**:785–794.
McGowen MR, Tsagkogeorga G, Álvarez-Carretero S, Dos Reis M, Struebig M, Deaville R, Jepson PD, Jarman S, Polanowski A, Morin PA, Rossiter SJ. 2020. Phylogenomic Resolution of the Cetacean Tree of Life Using Target Sequence Capture. *Syst Biol* **69**:479–501.
Miralles L, Oremus M, Silva MA, Planes S, Garcia-Vazquez E. 2016. Interspecific Hybridization in Pilot Whales and Asymmetric Genetic Introgression in Northern Globicephala melas under the Scenario of Global Warming. *PLoS One* **11**:e0160080.
Miyazaki N, Hirosaki Y, Kinuta T, Omura H. 1992. Osteological study of a hybrid between Tursiops truncatus and Grampus griseus. *Bull Natl Mus Nat Sci Ser B Bot* **18**:79–94.
Moodley Y, Westbury MV, Russo I-RM, Gopalakrishnan S, Rakotoarivelo A, Olsen R-A, Prost S, Tunstall T, Ryder OA, Dalén L, Bruford MW. 2020. Interspecific gene flow and the evolution of specialisation in black and white rhinoceros. *Mol Biol Evol*. doi:10.1093/molbev/msaa148
Moura AE, Kenny JG, Chaudhuri RR, Hughes MA. 2015. Phylogenomics of the killer whale indicates ecotype divergence in sympatry. *Heredity* **114**:48–55.
Moura AE, Shreves K, Pilot M, Andrews KR, Moore DM, Kishida T, Möller L, Natoli A, Gaspari S, McGowen M, Chen I, Gray H, Gore M, Culloch RM, Kiani MS, Willson MS,

bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Bulushi A, Collins T, Baldwin R, Willson A, Minton G, Ponnampalam L, Hoelzel AR. 2020. Phylogenomics of the genus *Tursiops* and closely related Delphininae reveals extensive reticulation among lineages and provides inference about eco-evolutionary drivers. *Mol Phylogenet Evol* **146**:106756.

Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* **32**:1749–1751.

Norris RD, Hull PM. 2012. The temporal dimension of marine speciation. *Evol Ecol* **26**:393–415.

Palumbi SR. 1994. Genetic divergence, reproductive isolation, and marine speciation. *Annu Rev Ecol Syst* **25**:547–572.

Pease JB, Hahn MW. 2015. Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Syst Biol* **64**:651–662.

Pease JB, Rosenzweig BK. 2018. Encoding Data Using Biological Principles: The Multisample Variant Format for Phylogenomics and Population Genomics. *IEEE/ACM Trans Comput Biol Bioinform* **15**:1231–1238.

Polyak VJ, Onac BP, Fornós JJ, Hay C, Asmerom Y, Dorale JA, Ginés J, Tuccimei P, Ginés A. 2018. A highly resolved record of relative sea level in the western Mediterranean Sea during the last interglacial period. *Nat Geosci* **11**:860–864.

Silva JM, Silva FJL, Sazima I. 2005. Two presumed interspecific hybrids in the genus *Stenella* (Delphinidae) in the Tropical West Atlantic. *Aquat Mamm* **31**:468.

Skovrind M, Castruita JAS, Haile J, Treadaway EC, Gopalakrishnan S, Westbury MV, Heide-Jørgensen MP, Szpak P, Lorenzen ED. 2019. Hybridization between two high Arctic cetaceans confirmed by genomic analysis. *Sci Rep* **9**:7729.

Slatkin M, Pollack JL. 2008. Subdivision in an ancestral species creates asymmetry in gene trees. *Mol Biol Evol* **25**:2241–2246.

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313.

Steeman ME, Hebsgaard MB, Fordyce RE, Ho SYW, Rabosky DL, Nielsen R, Rahbek C, Glenner H, Sørensen MV, Willerslev E. 2009. Radiation of extant cetaceans driven by restructuring of the oceans. *Syst Biol* **58**:573–585.

Stone G, Florez-Gonzalez L, Katona S. 1990. Whale migration record. *Nature* **346**:705–705.

Turelli M, Barton NH, Coyne JA. 2001. Theory and speciation. *Trends Ecol Evol* **16**:330–343.

Westbury MV, Hartmann S, Barlow A, Preick M, Ridush B, Nagel D, Rathgeber T, Ziegler R, Baryshnikov G, Sheng G, Ludwig A, Wiesel I, Dalen L, Bibi F, Werdelin L, Heller R, Hofreiter M. 2020. Hyena paleogenomes reveal a complex evolutionary history of cross-continental gene flow between spotted and cave hyena. *Science Advances* **6**:eaay0456.

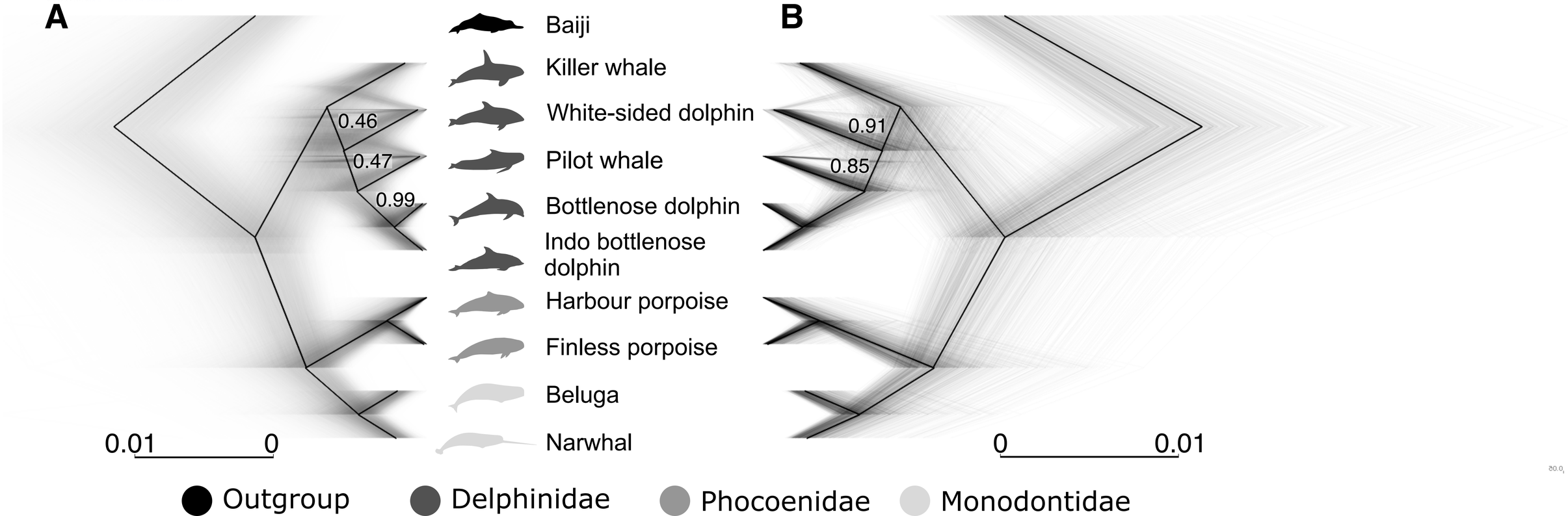
Westbury MV, Petersen B, Lorenzen ED. 2019. Genomic analyses reveal an absence of contemporary introgressive admixture between fin whales and blue whales, despite known hybrids. *PLoS One* **14**:e0222004.

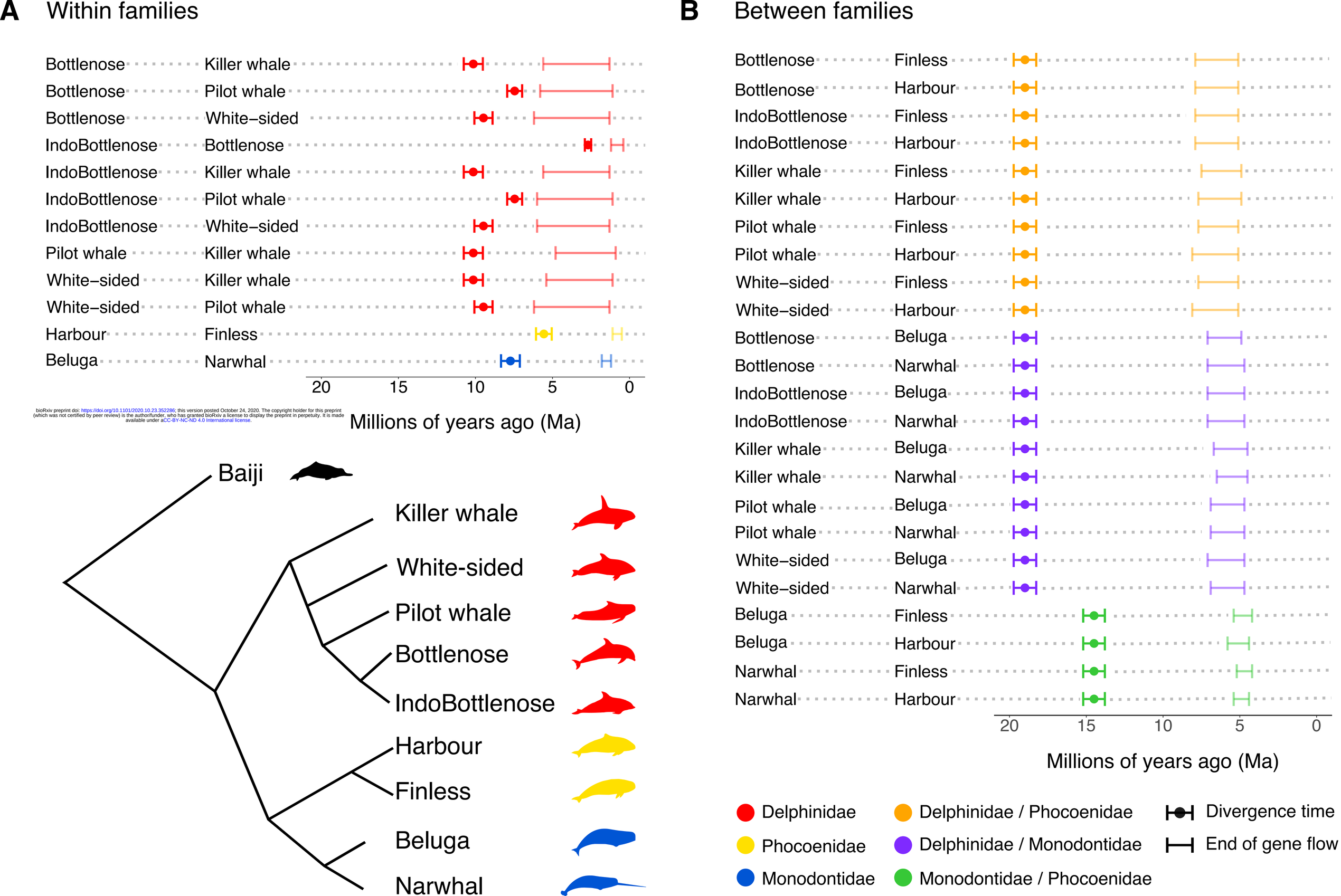
Williams TM. 1999. The evolution of cost efficient swimming in marine mammals: limits to energetic optimization. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences* **354**:193–201.

Willis PM, Crespi BJ, Dill LM, Baird RW, Hanson MB. 2004. Natural hybridization between Dall's porpoises (*Phocoenoides dalli*) and harbour porpoises (*Phocoena phocoena*). *Can J Zool* **82**:828–834.

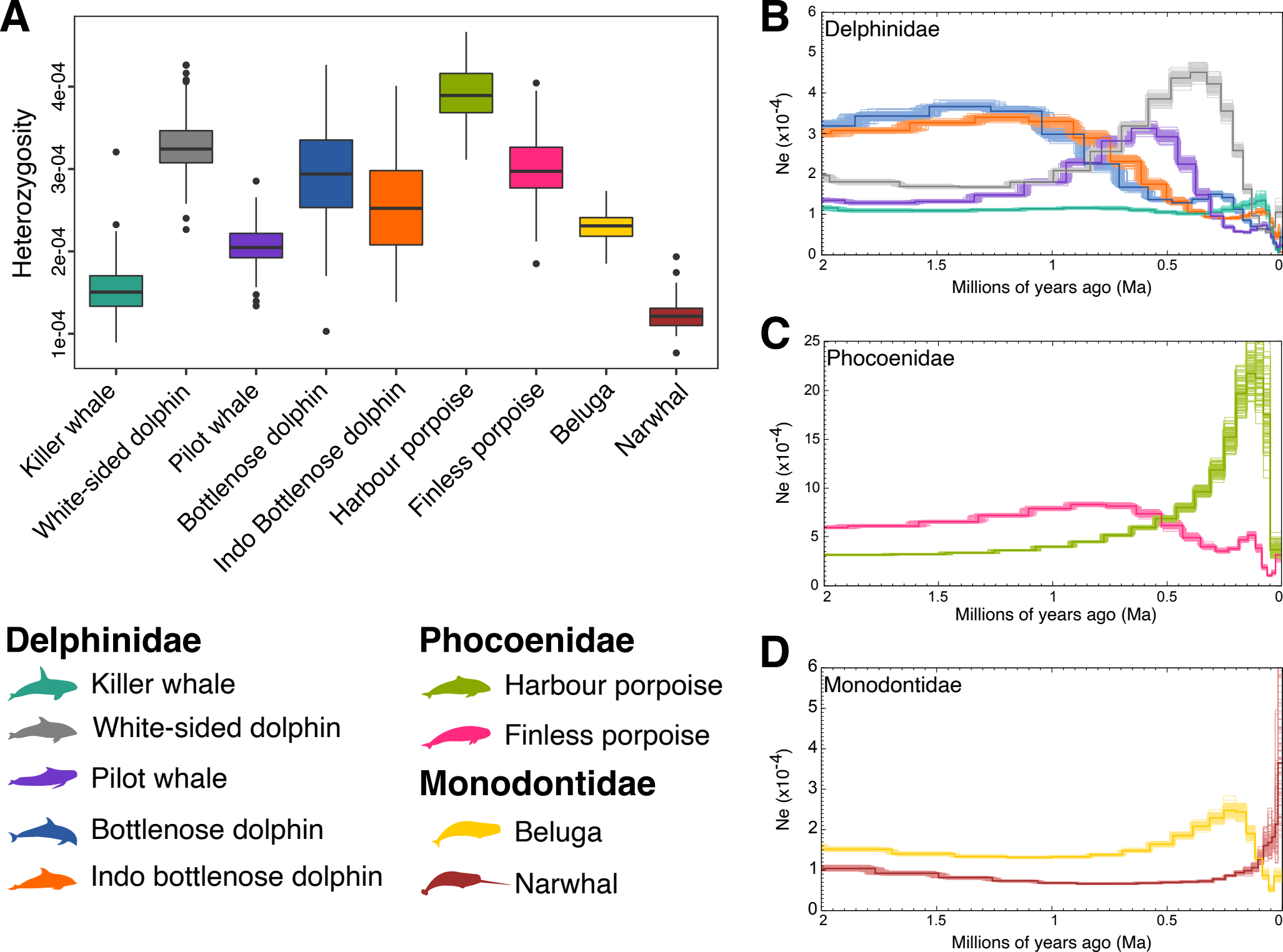
Zheng Y, Janke A. 2018. Gene flow analysis method, the D-statistic, is robust in a wide parameter space. *BMC Bioinformatics* **19**:10.

(which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.





bioRxiv preprint doi: https://doi.org/10.1101/2020.10.23.352286; this version posted October 24, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.


A
Heterozygosity
1e-04
2e-04
3e-04
4e-04
Killer whale
White-sided dolphin
Pilot whale
Bottlenose dolphin
Indo Bottlenose dolphin
Harbour porpoise
Finless porpoise
Beluga
Narwhal
B
Delphinidae
Ne (x10^-4)
Millions of years ago (Ma)
C
Phocoenidae
Ne (x10^-4)
Millions of years ago (Ma)
D
Monodontidae
Ne (x10^-4)
Millions of years ago (Ma)
Delphinidae
Killer whale
White-sided dolphin
Pilot whale
Bottlenose dolphin
Indo bottlenose dolphin
Phocoenidae
Harbour porpoise
Finless porpoise
Monodontidae
Beluga
Narwhal