GOETHE
UNIVERSITÄT
FRANKFURT AM MAIN

The application of baleen whale genomes in conservation and evolutionary research.

Dissertation
Zur Erlangung des Doktorgrades
der Naturwissenschaften

vorgelegt beim Fachbereich 15 Biowissenschaften
der Johann Wolfgang Goethe - Universität
in Frankfurt am Main

von
Magnus Wolf
aus Schwelm

Frankfurt (2023)
(D 30)

Vom Fachbereich 15 Biowissenschaften der
Johann Wolfgang Goethe - Universität als Dissertation
angenommen.

Dekan: Prof. Dr. Sven Klimpel

Gutachter: Prof. Dr. Axel Janke, Prof. Dr. Ingo Ebersberger

Datum der Disputation:

“*If I have seen further than others, it is by standing upon the shoulders of giants.*”

– Sir Isaac Newton–

For my little son Matheo.

Table of Contents / Inhaltsverzeichnis

1. Summary

1.1 Background

Baleen whales (Mysticeti) are a clade of highly adapted carnivorous marine mammals that can reach extremely large body sizes and feature characteristic keratinaceous baleen plates used for obligate filter feeding. Most baleen whales participate in seasonal migrations and many species are known to travel large distances between feeding and breeding grounds.

Due to their enormous size, often referred to as gigantism, baleen whales are assumed to have an extensive impact on marine ecosystems by e.g., enhancing nutrient recycling. Despite their large bodies, baleen whales experience unusually low rates of tumor development which promises helpful insights in medical research, provided that the responsible genes and molecular patterns are identified. Because of the high amounts of different products like e.g., oil that can be retrieved from catching a single individual and the historically high demand of these products, baleen whales were hunted extensively over a roughly 100 years lasting time period that depleted many of the respective whale stocks with so far unknown consequences for e.g. their molecular viability.

Over the last ~50 million years, baleen whales underwent multiple, remarkable morphological transitions that are considered to be related to major changes of ocean and climate dynamics within earth's history. Starting from land-dwelling, actively hunting artiodactyls, they shifted stepwise to today's fully aquatic filter feeding animals. So far, it is still debated which factors exactly facilitated these transitions and how such extensive macro-evolutionary changes happened on the micro-evolutionary and molecular level.

Our understanding of the long-term consequences of the industrial whaling era as well as of the exact nature of baleen whale evolution is still lacking, most likely impeded by their highly migratory behavior in an inaccessible open sea which inevitably results in logistical challenges. Furthermore, Cetaceans have a sparse fossil record, lacking many pieces especially in their early evolution. Analyzing molecular data promises new insights into these questions and could help mitigate the faced challenges in classical biological research.

In this dissertation, I will demonstrate the application of baleen whale genomes to tackle these open questions by using modern approaches of conservation and evolutionary genomics.

1.2 Realized studies

In **Publication 1** (Wolf et al. 2022, *Molecular Biology and Evolution*, Volume 39, Issue 5), I evaluated the impact of the industrial whaling era on the molecular viability of an Icelandic fin whale (*Balaenoptera physalus*) population. North Atlantic fin whales were subject to large-scaled hunting operations since the beginning of industrial whaling and a fist local over-exploitation was reached in 1904, with a so far unknown extent and unknown consequences for the population.

To assess these unknowns, the genomes of 51 fin whale individuals were sequenced, representing three temporally separated intervals in time, namely 1989, 2009 and 2018, which allowed measurements of potential changes over the last 30 years. Genomic data was sequenced using short read technology and linked short reads were used to construct a first reference genome assembly of the species. Demographic models based on the side frequency spectrum made it possible to assess the extent of the bottleneck imposed on the North-Atlantic fin whale during the whaling era. Furthermore, many aspects of the molecular viability of the population were analyzed, such as genome wide heterozygosity, inbreeding by the means of runs of homozygosity and mutational load by identifying potential deleterious mutations.

The results suggest a substantial drop in the effective population size but also an otherwise lack of manifestation in their genotypes. Levels of heterozygosity were well in the range of other whales and mammals, runs of homozygosity were short which indicates frequent outcrossing, and no excess of deleterious mutations was found, speaking for no apparent fitness reduction of the population. These results were put into context by comparisons to available single genomes of other baleen whale species which suggested that other more threatened species like the blue whale (*Balaenoptera musculus*) and the North Atlantic right whale (*Eubalaena glacialis*) may be more affected in their molecular viability as indicated by the presence of long runs of homozygosity and higher amounts of potential deleterious mutations in otherwise more heterozygous genomes. Eventually, these results indicated that genetic diversity alone may not be directly informative for the fitness of a whale population and that this kind of analysis needs to be complemented with other measures.

In **Publication 2** (Wolf et al. 2023, *BMC Biology*,Volume 21, Issue 79), I constructed the genome of the pygmy right whale (*Caperea marginata*) and tested its potential in phylogenetics and cancer research. The species is the smallest among baleen whales, occurs circumpolar in Antarctic waters, and represents the last representative of an otherwise extinct family of whales (Cetotheriidae). It is assumed that after the split from the Cetotheriidae, the rorquals

(Balaenopteridae) quickly evolved into different lineages which may have resulted in the unclear relationships and contested hypothesis about their early divergence in the past. Furthermore, the unique small size of the pygmy right whale might facilitate the search for genes related to size and cancer resistance when compared to molecular data of gigantic relatives.

Phylogenomic analyses using fragments of a whole-genome alignment featuring nearly all extant baleen whales, allowed the revision of the complex evolutionary relationships of rorquals by quantifying and characterizing the amounts of conflicts in early diverging branches. These relationships were further used to identify phylogenetically independent pairs of baleen whales with a maximum of diverging body size differences to compare rates of positive selection between their genomes.

The results suggest nearly evenly distributed frequencies of alternative topologies which supports the representation of the early divergence of rorquals as a hard polytomy with high amounts of introgression and incomplete lineage sorting. Within the set of available genomic data, three independent pairs of baleen whales with diverging body sizes were found and comparisons of positive selection rates resulted in many potentially body size and cancer related genes. The lack of conserved selection patterns, however, suggest a more convergent evolution of size and cancer resistance like previously discussed in paleontology.

In **Manuscript 1** (Wolf et al. submitted, *Molecular Ecology*, manuscript ID: MEC-23-0421), I measured rates of genome-wide isolation and divergence between two populations of the northern-hemisphere blue whale (*Balaenoptera musculus musculus*). Due to their inaccessibility, many whales are expected to include so far uncharacterized isolated populations or subspecies and the lack of classifications may impede further conservation management. Within the northern hemisphere, blue whales exist in both the Atlantic and Pacific Ocean, separated by major land masses and already known acoustic differences, suggesting potential accumulated differences. Furthermore, Publication 1 already indicated potential consequences of the whaling era on the molecular viability of the blue whale and assessing the degree in both populations might further help and support their conservation efforts.

Genomes of 14 North-Pacific blue whales were sequenced using short read data and complemented with publicly available 11 genomes of North-Atlantic blue whales. To contextualize the results with an uncontested subspecies, six genomes of the well-established Indo-Australian pygmy blue whale (*Balaenoptera musculus brevicauda*) were sequenced as well. Furthermore, to estimate found differences compared to a closely related species, one genome of the sei whale (*Balaenoptera borealis*) was sequenced and added to the dataset together with two publicly available genomes.

Population genetic and gene flow analyses showed clearly separated and well isolated populations in accordance with their assumed geographical distance. The genome-wide divergence, however, was low compared to other cetacean populations and to the next closely related sei whale species. Because this includes the morphologically different and well recognized pygmy blue whale subspecies, a proposal was made to equally categorize the two northern-hemisphere blue whale populations as subspecies. Furthermore, conservation genomic aspects like genetic diversity, neutral evolution tests and traces of inbreeding suggested, in accordance with Publication 1, a rather high impact of their depletion on the molecular viability of the species by the means of long runs of homozygosity and lack of rare alleles in otherwise highly heterozygous genomes.

1.3 Conclusion

The application of whole genome data using methods of conservation genetics allowed for a comprehensive estimation about the molecular viability of blue and fin whales as well as an assessment of the taxonomic status of northern-hemisphere blue whale populations. The rather different results between blue and fin whales underlines the importance of genomic monitoring of baleen whales because different species show rather different molecular consequences of their potentially varying depletions. Furthermore, as showcased for the northern-hemisphere blue whale, many important isolated populations of baleen whales may still be unknown to conservation management and genome-wide comparisons will most likely contribute to overcome this under-classification problem.

The application of whole genome data in evolutionary research allowed the characterization of the complex patterns of molecular conflicts within baleen whales and especially rorquals that will contribute to the still rather unclear understanding of their evolution. The here found molecular support for the idea of convergent evolution of gigantism in whales will further guide the search for molecular patterns responsible for Peto’s paradox.

2. Zusammenfassung

Anwendung genomischer Daten von Bartenwalen in Naturschutz- und Evolutionsforschung.

2.1 Hintergrund

Bartenwale (Mysticeti) sind eine hochspezialisierte Gruppe karnivorer, mariner Säugetiere, die extreme Körpergrößen erreichen können und die mit speziellen, aus Kreatin bestehenden Barten, tierisches Plankton oder Fische aus dem Wasser filtern. Die meisten Bartenwale unternehmen saisonale Wanderungen, bei denen sie zum Teil große Distanzen überbrücken, um zwischen Gewässern für die Nahrungsaufnahme und für die Fortpflanzung hin und her zu wechseln.

Wegen ihrer enormen Größe, die oft auch als Gigantismus bezeichnet wird, wird davon ausgegangen, dass Bartenwale einen weitreichenden Einfluss auf marine Ökosysteme haben, da ihre große Aufnahme und Verwertung von Biomasse große Mengen an Nährstoffen freisetzt. Trotz ihrer Größe, die viele Zellen und Zellteilungen mit sich bringt, leiden Bartenwale vergleichsweise selten an Tumorerkrankungen. Die Entschlüsselung dieses Phänomens könnte daher zu Fortschritten der medizinischen Forschung beitragen, sofern die entsprechenden genetischen und molekularen Muster identifiziert werden können, die für diese Resistenz verantwortlich sind. Die großen Körper der Bartenwale bedeuteten auch, dass der Fang eines einzigen Individuums große Mengen an körpereigenem Tran eintrug. Aufgrund der historisch hohen Nachfrage nach daraus gewonnenen Produkten wie etwa Öl, führte dies zu einer etwa 100 Jahre andauernden Epoche des industriellen Walfangs, die viele Populationen stark schrumpfen ließ. Mögliche Folgen dieser Ausbeutung auf zum Beispiel die Genetik der Bartenwale sind bisher kaum erforscht.

In den letzten ~50 Millionen Jahren haben Wale (Cetacea) mehrere tiefgreifende morphologische Veränderungen durchlaufen, die vermutlich mit ebenso weitreichenden Veränderungen der Meeresströmungen und Klimaveränderungen innerhalb der Weltgeschichte zusammenhängen. Dabei wird davon ausgegangen, dass Bartenwale von ehemals landlebenden, aktiv jagenden Säugetieren abstammen, die in die Vorfahren der heutigen

Paarhufer (Artiodactyla) eingeordnet werden. Welche erdgeschichtlichen Faktoren diese weitreichenden Anpassungen begünstigt haben, ist noch immer umstritten, vor allem aber, wie sich diese makro-evolutionäre Veränderung auf der mikro-evolutionären Ebene und der molekularen Ebene vollzogen haben.

Dass unser Verständnis von sowohl dem Einfluss des industriellen Walfangs als auch von dem genauen Ablauf der Evolution der Bartenwale so ungenau ist, liegt vermutlich an der hohen Mobilität dieser Tiere und der schlechten Zugänglichkeit ihrer Lebensräume im offenen Meer. Dies führte in der Vergangenheit zu enormen logistischen Herausforderungen, die die Untersuchung dieser Tiere zum Teil unmöglich machte. Des Weiteren sind Fossilienfunde von frühen Walen selten und decken ihre Geschichte nur ungleichmäßig ab. In diesem Kontext könnte die Analyse von molekularen Daten deutlich mehr zum Verständnis dieser zwei Themen beitragen als die bisher angewandten klassischen Methoden der Biologie.

In dieser Dissertation wird gezeigt, wie die Anwendung von genomischen Daten und modernen Methoden der Naturschutz- und Evolutionsgenetik zum Verständnis der Biologie der Bartenwal beitragen kann, mit besonderem Schwerpunkt auf deren Schutz und Evolution.

2.2 Durchgeführte Studien

In **Publikation 1** (Wolf et al. 2022, *Molecular Biology and Evolution*, Volume 39, Issue 5), habe ich den Einfluss des industriellen Walfangs auf die molekularen Eigenschaften einer isländischen Finnwalpopulation (*Balaenoptera physalus*) untersucht. Finnwale im Nordatlantik waren schon seit Beginn industrieller Fangfahrten ein begehrtes Ziel und eine Überfischung der Bestände wurde bereits 1904 dokumentiert, mit bisher unbekanntem Ausmaß und unbekannten molekularen Konsequenzen für die Population.

Um diese besser einschätzen zu können, wurden die kompletten Genome von 51 Finnwalindividuen aus drei zeitlich versetzten Jahren sequenziert, um mögliche zeitliche Veränderungen innerhalb der letzten 30 Jahre feststellen zu können. Die genomischen Daten wurden mittels Hochdurchsatz-Methoden erstellt und eine Genom-Assemblierung der Art zu Referenzzwecken anhand verlinkter Sequenzfragmente durchgeführt. Demographische Modelle, die auf dem Spektrum von Allelfrequenzen basieren, wurden verwendet, um das Ausmaß der etwa 100 Jahre zurückliegenden Bejagung des Nordatlantischen Finnwals zu ermitteln. Des Weiteren wurden verschiedene molekulare Aspekte untersucht, die vom Rückgang der Population betroffen sein könnten. Dazu gehören die genetische Diversität, die durch genomweite Heterozygotie bestimmt wurde, die Länge und Frequenz homozygoter

Abschnitte, die den Einfluss von Inzucht widerspiegeln und zuletzt die Anzahl potenziell schädlicher Mutationen, die Aufschluss über den Fitnesszustand einer Population geben können.

Die Ergebnisse weisen zwar auf eine starke Reduktion der Populationsgröße innerhalb des industriellen Walfangs hin, aber auch auf einen ansonsten eher geringen molekularen Einfluss der Bejagung. Die genetische Diversität der Population ist vergleichsweise hoch und lange homozygote Abschnitte oder ein Überschuss an schädlichen Mutationen fehlen. Um diese Ergebnisse im Kontext anderer Bartenwale zu interpretieren, wurden die entsprechenden Werte mit öffentlichen Genomen anderer Arten verglichen. Dabei zeigte sich, dass gerade die stärker bedrohten Walarten wie etwa der Blauwal (*Balaenoptera musculus*) und der Atlantische Nordkaper (*Eubalaena glacialis*), eine besonders hohe Heterozygotie, aber auch besonders viele und lange homozygote Abschnitte und eine erhöhte Anzahl schädlicher Mutationen aufweisen, was für eine erhöhte Inzuchtrate und eventuelle Fitnesseinflüsse spricht. Die Ergebnisse zeigen aber auch, dass die genetische Diversität nicht unbedingt den direkten Fitnesszustand einer Walpopulation widerspiegelt und dass bei solchen Analysen immer auch andere Faktoren untersucht werden sollten.

In **Publikation 2** (Wolf et al. 2023, *BMC Biology*,Volume 21, Issue 79), habe ich das Genom des Zwergglattwals (*Caperea marginata*) rekonstruiert und das Potential des neuen Referenzgenoms in der Phylogenetik und Tumorforschung getestet. Der Zwergglattwal ist der kleinste Vertreter der Bartenwale und der letzte Vertreter einer ansonsten ausgestorbenen Familie (Cetotheriidae). Es wird angenommen, dass sich die Furchenwale (Balaenopteridae) nach ihrer Trennung von den Cetotheriidae, schnell in mehrere Linien aufgespalten haben, was ihre evolutionären Beziehungen unklar macht und zu verschiedenen, oft widersprüchlichen Hypothesen über ihre genaue Evolution führte. Die geringe Körpergröße des Zwergglattwals könnte weiterhin die Suche nach Genen, die für den Walgigantismus und die damit verbundene Krebsresistenz verantwortlich sind, erleichtern, wenn die entsprechenden Sequenzen mit denen seiner gigantischen Verwandten verglichen werden.

Eine phylogenomische Analyse auf der Basis von Fragmenten eines Alignments ganzer Genome erlaubte es, die Verwandschaftsbeziehungen der Bartenwale und insbesondere die komplexen evolutionären Konflikte innerhalb der Furchenwale genauestens zu charakterisieren und zu quantifizieren. Die beschriebenen Beziehungen wurden dann verwendet, um phylogenetisch unabhängige Paare innerhalb der Bartenwale zu identifizieren, die die insgesamt größten Unterschiede in ihrer Körpergröße aufwiesen. Die Genome dieser Paare wurden dann auf Anzeichen einer möglichen positiven Selektion untersucht.

Die Ergebnisse deuten darauf hin, dass alternative Topologien nahezu gleich häufig auftreten und dass die frühe Aufspaltung der Furchenwale am besten als echte Polytomie beschrieben werden kann, die mit einem hohen Grad an Introgression und Transspezies-Polymorphismus einherging. Weiterhin wurden drei unabhängige Paare mit großen Differenzen in ihrer Körpergröße identifiziert, deren genomweiter Vergleich der Selektionsraten zu einer hohen Anzahl an Genkandidaten führte, die für die charakteristische Tumorresistenz der Wale verantwortlich sein könnten. Allerdings wurde auch ein Mangel an Genen festgestellt, die in allen großen Walen gleichermaßen positiv selektiert werden, was eine in der Paläontologie bereits diskutierte Idee des konvergent entstandenen Walgigantismus unterstützt.

In **Manuskript 1** (Wolf et al. submitted, *Molecular Ecology*, Manuskript Nr.: MEC-23-0421), habe ich die genomweite Isolation und Divergenz zwischen zwei Populationen der nördlichen Blauwal-Unterart *Balaenoptera musculus musculus* untersucht. Wegen ihrer schlechten Zugänglichkeit wird bei Walen allgemein davon ausgegangen, dass es noch viele unbekannte isolierte Populationen oder Unterarten gibt und dieser Mangel an taxonomischer Klassifikation könnte den Schutz der Populationen behindern. In der nördlichen Hemisphäre gibt es Blauwale sowohl im Atlantik als auch Pazifik, getrennt von großen Landmassen und mit schon bekannten akustischen Unterschieden, die für eine bereits bestehende Ansammlung von Differenzen sprechen könnte. Publikation 1 hat außerdem gezeigt, dass der industrielle Walfang deutliche molekulare Konsequenzen für den Blauwal gehabt haben könnte und die Einschätzung dieser genetischen Faktoren zum Schutz des Blauwals beitragen könnte.

In diesem Projekt wurden 14 Genome von Blauwalen aus dem Nordpazifik mittels kurz-Fragment Methoden sequenziert und mit 11 öffentlichen Genomen nordatlantischer Blauwale verglichen. Um die gefundenen Unterschiede zu kontextualisieren, wurden zusätzlich sechs Genome der weithin anerkannten Unterart des indo-australischen Zwergblauwals (*Balaenoptera musculus brevicauda*) sequenziert. Außerdem wurde ein Genom des nahe verwandten Seiwals (*Balaenoptera borealis*) sequenziert und mit zwei bereits veröffentlichten Genomen in den Datensatz aufgenommen, um die Ergebnisse mit Unterschieden auf Artniveau zu vergleichen.

Die Ergebnisse zeigten eine klare Isolation beider Populationen, die zu der des Zwergblauwals vergleichbar war, allerdings zeigten alle drei Populationen auch eine geringe genetische Divergenz verglichen mit anderen Walpopulationen und dem nahe verwandten Seiwal. Da dies auch die genetische Divergenz des Zwergblauwals betraf, wurde eine gleiche Einordnung der nördlichen Populationen als verschiedene Unterarten vorgeschlagen. Zusätzlich wurden alle drei Populationen noch auf ihre molekularen Eigenschaften bezüglich ihrer Gefährdung hin untersucht und eine hohe genetische Diversität aber auch ein Mangel an seltenen Allelen und eine hohe Frequenz langer homozygote Abschnitte in allen drei Populationen festgestellt. Diese molekularen Eigenschaften könnten, wie in Publikation 1 bereits diskutiert, Folgen ihrer stärkeren Bejagung sein und eventuell die Fitness der Tiere langfristig beeinflussen.

2.3 Fazit

Die Anwendung genomischer Daten in Kombination mit Methoden der Naturschutzgenetik hat es möglich gemacht, die molekularen Auswirkungen des Walfangs auf Finn- und Blauwalpopulationen abzuschätzen, als auch den taxonomischen Status der nördlichen Blauwalpopulationen neu zu bewerten. Die deutlich unterschiedlichen Ausmaße dieser Konsequenzen zwischen Finnwalen und Blauwalen unterstreicht die Wichtigkeit solcher genomischen Beobachtungen von Bartenwal-Populationen, da verschiedene Populationen sehr unterschiedlich vom Walfang betroffen sein könnten und deswegen unterschiedlich starke molekulare Konsequenzen davon getragen haben könnten. Außerdem, wie anhand der nördlichen Blauwale gezeigt werden konnte, sind dem Naturschutz wahrscheinlich noch viele isolierte Populationen und vielleicht sogar Unterarten unbekannt und der Vergleich genomischer Daten birgt ein großes Potential, dieser Unterklassifizierung effektiv entgegenzuwirken.

Die Anwendung genomischer Daten in der Evolutionsforschung hat es möglich gemacht, die komplexen Muster molekularer Konflikte innerhalb der Bartenwal-Klade und besonders innerhalb der Furchenwale zu charakterisieren und zum bisher ungenauen Verständnis ihrer Evolution beizutragen. Insbesondere die molekularen Hinweise auf die konvergente Evolution des Walgigantismus können die zukünftige Suche nach den molekularen Mustern, welche für das Peto Paradoxon verantwortlich sind, entscheidend erleichtern.

3. General Introduction

3.1 Baleen whales

Baleen whales (Parvorder: Mysticeti) are a 24-36 million year old group of large carnivorous mammals possessing characteristic keratinaceous baleen plates which are used for obligate filter feeding (Figure 1A, Árnason et al. 2018; Deméré et al. 2008; Fordyce and Marx 2018).

3.1.1 General information

Apart of the filter feeding behavior, baleen whales are also known for their large size and they encompass a variety of body sizes, reaching between 5 to 6.5 meters and 3 to 3.5 metric tons on the small side (pygmy right whale, *Caperea marginata*) and 21 to 30 meters and 140 to 199 metric tons and the large side (blue whale, *Balaenoptera musculus*) (Branch, Abubaker et al. 2007; Kemper 2009; McClain et al. 2015). The unique method of filter feeding, and hence baleen like structures, are only known to exist much later in the evolutionary history of mysticetes. Yet, their large size was found to exist even in the earliest known fossils of mysticetes, such as the about eight meter long *Llanocetus denticrenatus* (Figure 1B, Lambert et al. 2017). Their generally large body size, sometimes referred to as gigantism, has been studied to understand the evolution of gigantism (Slater et al. 2017). Such studies were especially made in the context of cancer resistance because despite their large numbers of cells and cell divisions, baleen whales do not have higher rates of cancer development (See “Peto’s Paradox” in (Peto et al. 1975; Silva et al. 2023; Tollis et al. 2019).

Baleen whales occur in all major oceans, but prefer colder, more productive waters that benefit from deep-sea ocean dynamics like upwelling and frontal meandering (Figure 1C, Branch, Stafford et al. 2007). All mysticetes are highly mobile animals, known to migrate seasonally between more productive feeding and warmer or saver breeding grounds (Double et al. 2014; Silva et al. 2013). Furthermore, trans-oceanic migrations occur in many baleen whale species and all major oceans, with a record published in 2018 reporting a travel distance of more than 8,000 kilometers in a single lifetime (Hucke-Gaete et al. 2018). This high potential for dispersal makes whales an interesting target to study the processes behind sympatric speciation and niche differentiation, but also challenges research due to their inaccessibility in a vast open sea (Árnason et al. 2018; Foote and Morin 2015).

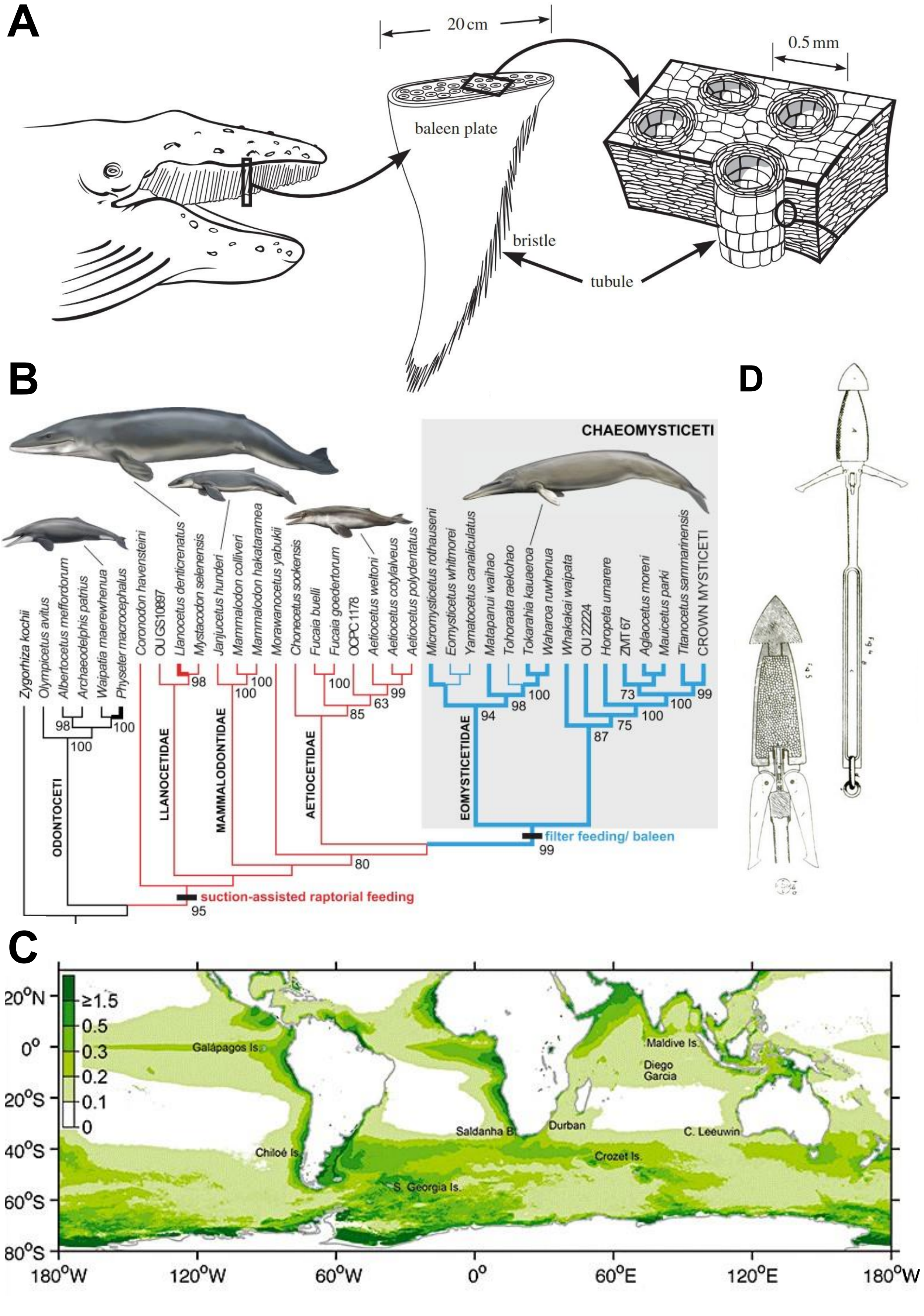
A
20 cm
0.5 mm
baleen plate
bristle
tubule
B
CHAEOMYSTICETI
Zygorhiza kochii
Olympicetus avitus
Albertocetus meffordorum
Archaeodelphis patrius
Waipatia maerewhenua
Physeter macrocephalus
Coronodon havensteini
OU GS10897
Llanocetus denticrenatus
Mystacodon selenensis
Janjucetus hunderi
Mammalodon colliveri
Mammalodon hakataramea
Morawanocetus yabukii
Chonecetus sookensis
Fucaia buelli
Fucaia goedertorum
OCPC 1178
Aetiocetus weltoni
Aetiocetus cotylalveus
Aetiocetus polydentatus
Micromysticetus rothauseni
Eomysticetus whitmorei
Yamatocetus canaliculatus
Matapanui waihao
Tohoraata raekohao
Tokarahia kauaeroa
Waharoa ruwhenua
Whakakai waipata
OU 22224
Horopeta umarere
ZMT 67
Aglaocetus moreni
Mauicetus parki
Titanocetus sammarinensis
CROWN MYSTICETI
98
100
100
98
100
100
100
99
63
85
100
94
98
73
100
99
75
100
87
80
99
95
ODONTOCETI
LLANOCETIDAE
MAMMALODONTIDAE
AETIOCETIDAE
EOMYSTICETIDAE
filter feeding/ baleen
suction-assisted raptorial feeding
C
20°N
0°
20°S
40°S
60°S
80°S
≥1.5
0.5
0.3
0.2
0.1
0
Galápagos Is.
Maldive Is.
Diego Garcia
Saldanha B.
Durban
C. Leeuwin
Chiloé Is.
Crozet Is.
S. Georgia Is.
180°W
120°W
60°W
0°
60°E
120°E
180°W
D
Fig 5
Fig 4

Figure 1 Summary of baleen whale specific traits, evolution and whaling. **A** Hierarchical morphology of humpback whale baleen, from the level of the whole baleen apparatus down to the horn tubules level (Szewciw et al. 2010). **B** Phylogenetic relationships of the ~8m long *Llanocetus denticrenatus* and the evolution of baleen whales traits like gigantism, suction-assisted feeding and baleen based filter feeding (Fordyce and Marx 2018). **C** Annual mean phytoplankton chlorophyll-*a* concentrations in $\mathrm{mg.m}^{-3}$ from SeaWiFS as a proxy for marine ecosystem productivity that limits suitable baleen whale habitats (Branch, Stafford et al. 2007). **D** Explosive "shell-harpoon", a technological key invention that facilitated the industrial whaling era, built by Svend Foyn (Tønnessen and Johnsen 1982).

Many baleen whales feature characteristic songs for communication (McDonald et al. 2006; Rekdahl et al. 2018) and they are expected to be highly intelligent given their complex social behavior and characteristic brain morphologies known from other animals considered as intelligent, including us humans (Butti et al. 2009; Wray et al. 2021).

3.1.2 Whaling and importance of conversation

All larger whale species were sought after by humans with a peak during the roughly 100 years lasting era of industrial whaling between the 1870's and 1970's (Tønnessen and Johnsen 1982). Fueled by the rise of new hunting technologies like explosive-harpoons and motorized ships together with an increased demand of whale-based products like oil, industrial whaling spread around the globe over all major oceans until their exploitation became insufficient due to dwindling catch rates and an increased competition by fossil based oil products (Figure 1D, Tønnessen and Johnsen 1982). Although this exploitation depleted many whale species and brought some to the brink of extinction, a complete whaling moratorium, established in 1982, resulted in a beginning recovery of many baleen whale stocks (Cooke 2018a, 2018b; IWC 1982).

However, the poor documentation of catch rates and difficulties to estimate pre- and post-whaling stocks sizes makes it challenging to evaluate the consequences for distinct baleen whale species. This also includes long-term consequences for the marine ecosystems inhabited by this species. Their enormous body sizes and hence large amount of consumed and digested biomass led to the expectation that filter feeding whales, like the large blue whales, may have contributed substantially to the pre-whaling marine ecosystem productivity by enhancing nutrient recycling (Savoca et al. 2021), in particular iron turnover through feces (Ratnarajah et al. 2016). Furthermore, in case of death, whale carcasses often sink to the deep-sea floor, creating a sudden enormous source of nutrients that is discussed to substantially contribute to the deep-sea ecosystem diversity (See "whale fall hypothesis", Smith et al. 2015).

Because of the unknowns and challenges regarding their conservation as well as other poorly understood parts of their biology, e.g. the evolution towards gigantism and hence cancer resistance, genetic data, especially genomic data, represents a promising alternative to tackle many of these questions. Outlining these opportunities for conservation and evolutionary research will be the main focus of this dissertation.

3.2 Age of genomics

The study of genomes in biology, known as the research field "Genomics", was coined by Dr. Thomas H. Roderick (Yadav 2007) while establishing the equally named scientific journal and describes the analyses of preferably whole genome data to answer biological questions. The field was pioneered by scientists working on DNA sequencing methods (Pareek et al. 2011; Ronaghi 2001; Sanger and Coulson 1975) and consortia established to sequence the first entire genomes from e.g. yeast, human or mouse (Goffeau et al. 1996; Venter et al. 2001; Waterston et al. 2002). These scientific achievements fueled the rise of new technologies leading to a steep decline in sequencing costs in parallel with steeply increasing computational resources that allowed processing the vast amounts of data (Wetterstrand 2021). Due to these developments, whole-genome sequences are now a fundamental part in biological research and are used to address various open questions in all kinds of related fields, such as conservation and evolution.

3.3 Conservation genomics

Conservation genomics is a loosely defined field of research that provides new information and ideas for conservation efforts using whole genome data. In recent years, this term has become a "*buzzword*" and largely refers to the already existing subfield of conservation genetics which derived from population genetics. The general goal is to understand the molecular dynamics within or between threatened populations in order to propose beneficial conservation efforts that could help to avoid extinction. In practice, two main applications have found broad interest in the scientific community, namely the assessment of negative genetic consequences for populations brought to or remaining at low numbers (Foote et al. 2021; Seth et al. 2021; van der Valk, Díez-Del-Molino et al. 2019) and the assessment of genetic exchange, population structures and genetic differentiation in order to define precise management units (Andrews et al. 2018; Attard et al. 2018; Walters and Schwartz 2021). However, their interdependent nature usually encourages addressing both directions in a combined analysis.

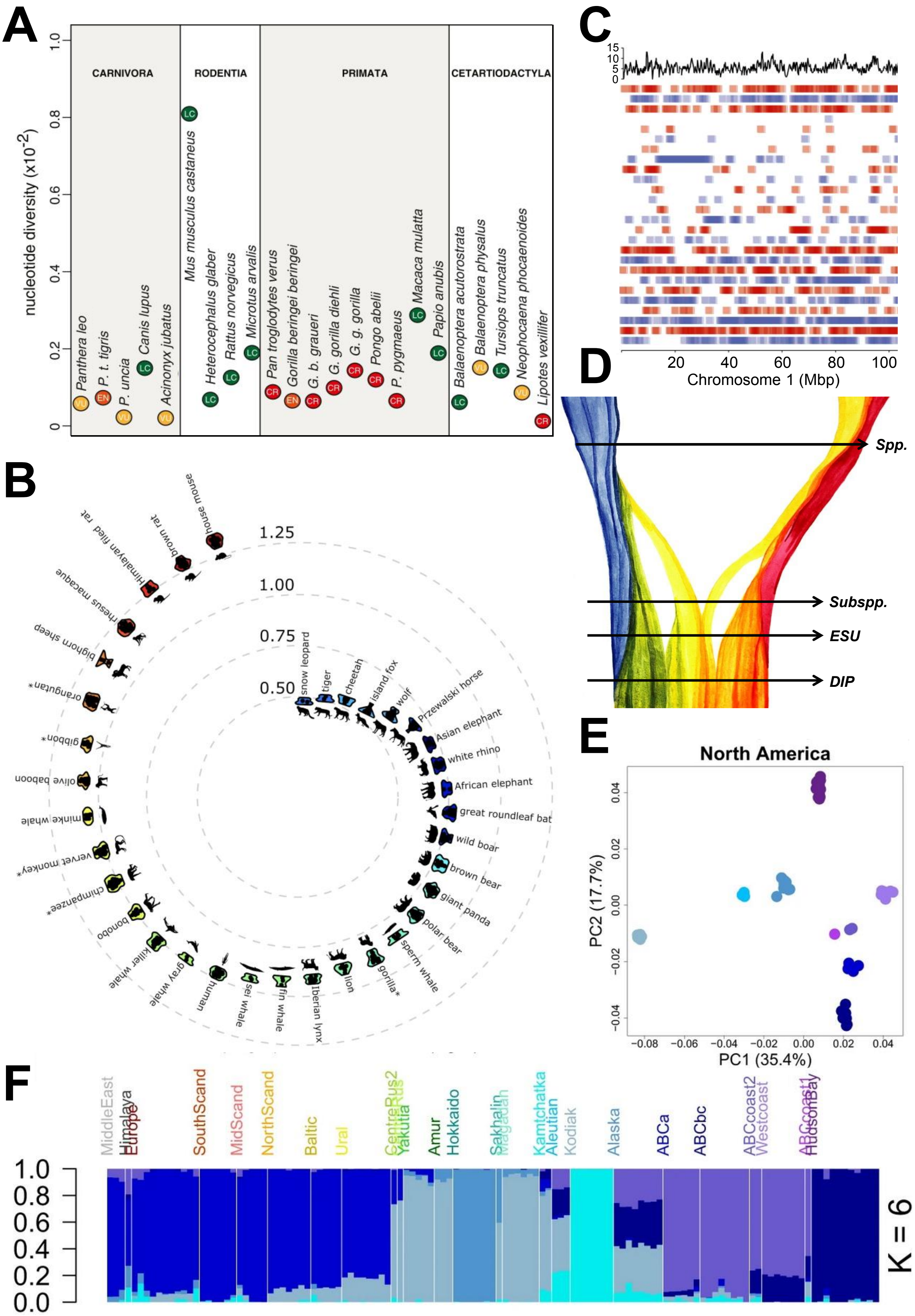
A
nucleotide diversity (x10^{-2})
0
0.2
0.4
0.6
0.8
1.0
CARNIVORA
RODENTIA
PRIMATA
CETARTIODACTYLA
Panthera leo
P. t. tigris
P. uncia
Canis lupus
Acinonyx jubatus
Mus musculus castaneus
Heterocephalus glaber
Rattus norvegicus
Microtus arvalis
Pan troglodytes verus
Gorilla beringei beringei
G. b. graueri
G. gorilla diehli
G. g. gorilla
Pongo abelii
P. pygmaeus
Macaca mulatta
Papio anubis
Balaenoptera acutorostrata
Balaenoptera physalus
Tursiops truncatus
Neophocaena phocaenoides
Lipotes vexillifer
VU
EN
LC
CR
B
1.25
1.00
0.75
0.50
snow leopard
tiger
cheetah
island fox
wolf
Przewalski horse
Asian elephant
white rhino
African elephant
great roundleaf bat
wild boar
brown bear
giant panda
polar bear
sperm whale
gorilla*
lion
Iberian lynx
fin whale
sei whale
human
gray whale
killer whale
bonobo
chimpanzee*
vervet monkey*
minke whale
olive baboon
gibbon*
orangutan*
bighorn sheep
rhesus macaque
Himalayan field rat
brown rat
house mouse
C
15
10
5
0
20
40
60
80
100
Chromosome 1 (Mbp)
D
Spp.
Subspp.
ESU
DIP
E
North America
PC2 (17.7%)
0.04
0.02
0.00
−0.02
−0.04
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
PC1 (35.4%)
F
MiddleEast
Himalaya
Europe
SouthScand
MidScand
NorthScand
Baltic
Ural
CentreRus2
CentreRus
Yakutia
Amur
Hokkaido
Sakhalin
Magadan
Kamtchatka
Aleutian
Kodiak
Alaska
ABCa
ABCbc
ABCcoast2
Westcoast
ABCcoast1
HudsonBay
1.0
0.8
0.6
0.4
0.2
0.0
K = 6

Figure 2 Applications of conservation genomic analyses. **A** Genome-wide nucleotide diversity is a poor predictor of IUCN's Red List status (Teixeira and Huber 2021). Colors indicate IUCN status e.g. least concern, vulnerable and endangered. **B** Genetic load over different species, here depicted as the average GERP-score of the derived allele for each individual within a species (van der Valk, Manuel et al. 2019). **C** Density of runs of homozygosity (ROH) along a chromosome of all ROH > 0.3 Mb in 26 globally sampled killer whale genomes (Foote et al. 2021). The black line indicates total counts, colored bars alter in red and blue and depict ROHs per individual. **D** Depiction of the divergence of lineages with four different hierarchical conservation management units, namely Demographically Independent Population (DIP), Evolutionarily Significant Unit (ESU), subspecies and species, defined by varying degrees of isolation and divergence (Taylor, Perrin et al. 2017). **E-F** Different population genetic analyses that indicate population structure and admixture in brown bears (Jong et al. 2023).

3.3.1 Genetic diversity

A main point of interest in the estimation of the "molecular viability" of a population is the genetic diversity and its implication for fitness. The term can be differentiated into the genetic diversity of an individual, like the amount of heterozygous sites within a genome and the genetic diversity of a population, here regarded as the genetic variation among a group of individuals. Genetic diversity is thought to affect both the adaptive and deleterious potential of segregating genetic variances within a population. While mutations tend to increase the genetic diversity of a population, genetic drift appears to reduce genetic diversity (Teixeira and Huber 2021). A reduced genetic diversity is often associated with a reduced adaptive potential (Spielman et al. 2004) or an increased deleterious potential (See Charlesworth and Willis 2009).

Adaptive potential is defined as the amount of additive genetic variation for adaptive traits between or among populations (Booy et al. 2000; Funk et al. 2019). In most cases, this potential is estimated based on the neutral genetic diversity of a population (Teixeira and Huber 2021), although the relationship between the two measures remains poorly understood and even questionable (Figure 2A, (Teixeira and Huber 2021). The deleterious potential on the other hand, goes along with the so-called genetic load and accounts for the number of mutations that could negatively affect the fitness of a population if becoming fixated (Figure 2B, Bertorelle et al. 2022; van der Valk, Manuel et al. 2019). Inbreeding, enhanced by small population sizes, plays an important role in this case, because mating between closely related individuals leads to larger runs of homozygous segments (ROH) of identity-by-descent that expose previously recessive deleterious mutations (Figure 2C, Charlesworth and Willis 2009; Foote et al. 2021). In extreme cases, the interrelationship between population size, genetic diversity, inbreeding and genetic load could lead to a down-spiraling process that continuously reduces fitness and population sizes, termed "inbreeding depression", "mutational meltdown" or "extinction vortex", depending on the factor identified as the main driver of this process (Charlesworth and Willis 2009; Fagan and Holmes 2006; Gabriel et al. 1993; Teixeira and Huber 2021). This process is inversely influenced by outbreeding with unrelated individuals which for example breaks up larger runs of homozygosity and reduces genetic load (Kim et al. 2018). Due to this positive effect of outbreeding on the fitness of a threatened population, the amount of genetic exchange represents an important measure of its own when evaluating the robustness of a population against severe depletion (Teixeira and Huber 2021).

3.3.2 Genetic exchange

Gene flow, here defined as the transfer of genetic material between two populations, is largely consistent with migration from one population to another and determines the independence and hence isolation of their genetic variances (Slarkin 1985). In conservation management, a main goal is to prevent isolation of a population that would otherwise drastically decrease the effective amount of genetic material in a population also known as gene pool (Frankham et al. 2017). Conservation strategies often aim to either ensure connectivity between fragmented populations or to protect large enough areas that facilitate a sufficient gene pool on its own (Frankham et al. 2017).

Therefore, the identification and evaluation of gene flow signals was always an important aspect of population genetics and hence conservation genetics and later conservation genomics (Figure 2E-F, Jong et al. 2023; Kozakiewicz et al. 2019; Slarkin 1985). The benefit of whole genome sequences in this instance is its ability to provide comprehensive information about the extent of genetic exchange that would not be possible with traditional marker sequences given that they can be affected rather differently by ongoing gene flow between two groups (Petit and Excoffier 2009). In case of an extensive isolation, populations start to accumulate specific genetic variances over time, a process that is measured in its genomic divergence and is largely speed up in small populations due to genetic drift, the randomized change of frequencies of genetic variances (Masel 2011; Palumbi 1994). Inevitably, these differences will manifest in different phenotypes, resulting in an increasing incompatibility between both populations and a beginning speciation process (Figure 2D, (Taylor, Perrin et al. 2017; Waples and Gaggiotti 2006). So-called genetic rescue programs, usually aiming to enhance gene flow and beneficial outcrossing, were already reported to be problematic because of either incompatibility between individuals or the harmful effects of recessive deleterious mutations in a highly inbred population (See list in: (Frankham et al. 2017), a phenomenon also regarded as outbreeding depression (Frankham et al. 2017; Ouborg et al. 2010). Population genomic comparisons can be applied in this instance to ensure reintroductions of more similar populations instead of individuals from a population that already accumulated considerable amounts of differences on their way towards speciation.

Finally, whole-genome sequences can also help to assess taxonomic uncertainties and identify previously unknown and potentially threatened conservation units like e.g. subspecies or species that may represent independently evolving gene pools. This so-called under-classification error might also lead to underrepresentation in received conservation efforts because unnamed conservation units are usually less likely to receive attention and hence protection (Taylor, Perrin et al. 2017). The genomic assessments made to identify independent taxonomic units usually aim to identify isolation and divergence, of which the latter arguably bears similarities to evolutionary research that also aims to identify genetic differences in the context of evolution.

3.4 Evolutionary genomics

Evolutionary genetics, the study of evolution based on molecular differences, was, equally to the study of conservation genetics, recently expanded by the use of whole genome data, resulting in the again loosely defined field of evolutionary genomics. A dominant goal in this field remains the reconstruction of evolution by the means of phylogenetic systematics (Hennig 1965), but also the study of the underlying mechanisms of evolution and the evolution of genomes themselves fall under this broad term too.

3.4.1 Phylogenomics

The construction of tree-like graphs to depict the evolutionary past of organisms has been done since Charles Darwin proposed his ideas of evolution (Figure 3A). Today, differences in the molecular components of life, such as DNA, RNA and peptides are used to study the evolutionary past and such analyses ultimately result in a hypothesis represented as a phylogenetic tree. A major cornerstone of this type of analysis is the assumption that similar molecular sequences represent homologous traits passed down by a common ancestor (Knoop and Müller 2009). Following this assumption, the amount of similarity is roughly informative about the evolutionary relationship or distance between two sequences and hence their originating organisms.

Over the years, many theoretical developments have revolutionized this discipline of reconstructing phylogenetic trees which eventually led to the idea to construct a “Tree of life” (Figure 3B). Technical advances were for example made by improving methods of sequence alignments that align similar and potential homologous information (Nakamura et al. 2018; Needleman and Wunsch 1970). Another point was the progressing understanding of the stochastics of evolutionary change that resulted in increasingly refined evolutionary models (Jukes and Cantor 1969; Shapiro et al. 2006). Methods of deciphering between genes descended from a common ancestor by speciation (orthologs) or gene duplication (paralogs) largely removed otherwise abundant artifacts, although their identification remains a challenging task with ongoing improvements (Linard et al. 2021). Tree construction algorithms were developed to infer a phylogeny on the basis of similarity, parsimony, maximum likelihood and Bayesian statistics (Felsenstein 1981; Gascuel 1997; Ronquist and Huelsenbeck 2003). At last, methods of tree evaluation were established to retrospectively analyze the support of a certain tree or specific branch (Felsenstein 1985).

Alongside these theoretical milestones, technical progress in both sequencing methods and computational resources happened as described above. Together, this enabled the construction of trees based on more taxa and more sequences eventually leading to what is now known as "phylogenomics", a term used to emphasize the large amount of incorporated data, preferably whole genome data (Bleidorn 2017). Different types of orthologous data were already used in phylogenomic reconstructions like whole genome alignments, sets of orthologous genes, transposable elements, and mitochondrial or chloroplast sequences with varying rates of evolutionary change and varying amounts of information and hence different capabilities to resolve different periods of the evolutionary past (Árnason et al. 2018; Janke et al. 1994; Lammers et al. 2019; McGowen et al. 2020; Wicke et al. 2014). No matter the type of data, the goal in phylogenetic reconstruction is to find the overall best tree-hypothesis. To get to this, single genetic variances, called alleles, are usually handled as independent evolutionary traits and they might, depending on their frequency, support a certain branching pattern also called topology. The easiest way to get an overall hypothesis is then to construct the most favored topology among all alleles by concatenating all sequence alignments into a single, matrix-like structure before running a tree-construction algorithm on this total set of sequences, exaggeratedly called "supermatrix" (Figure 3D, (Delsuc et al. 2005). Since different regions of the genome, here called loci, may support different topologies in their allelic patterns, it is sometimes also beneficial to construct smaller trees per locus and instead try to find a consensus tree between resulting set of trees (Kubatko and Degnan 2007), exaggeratedly called "supertree" (Figure 3D, Bininda-Emonds 2004). Similar to tree construction in general, many different approaches were developed over the years to find the best consensus tree, reaching from simply greedy consensus approaches (Gordon 1986) to more elaborate techniques like

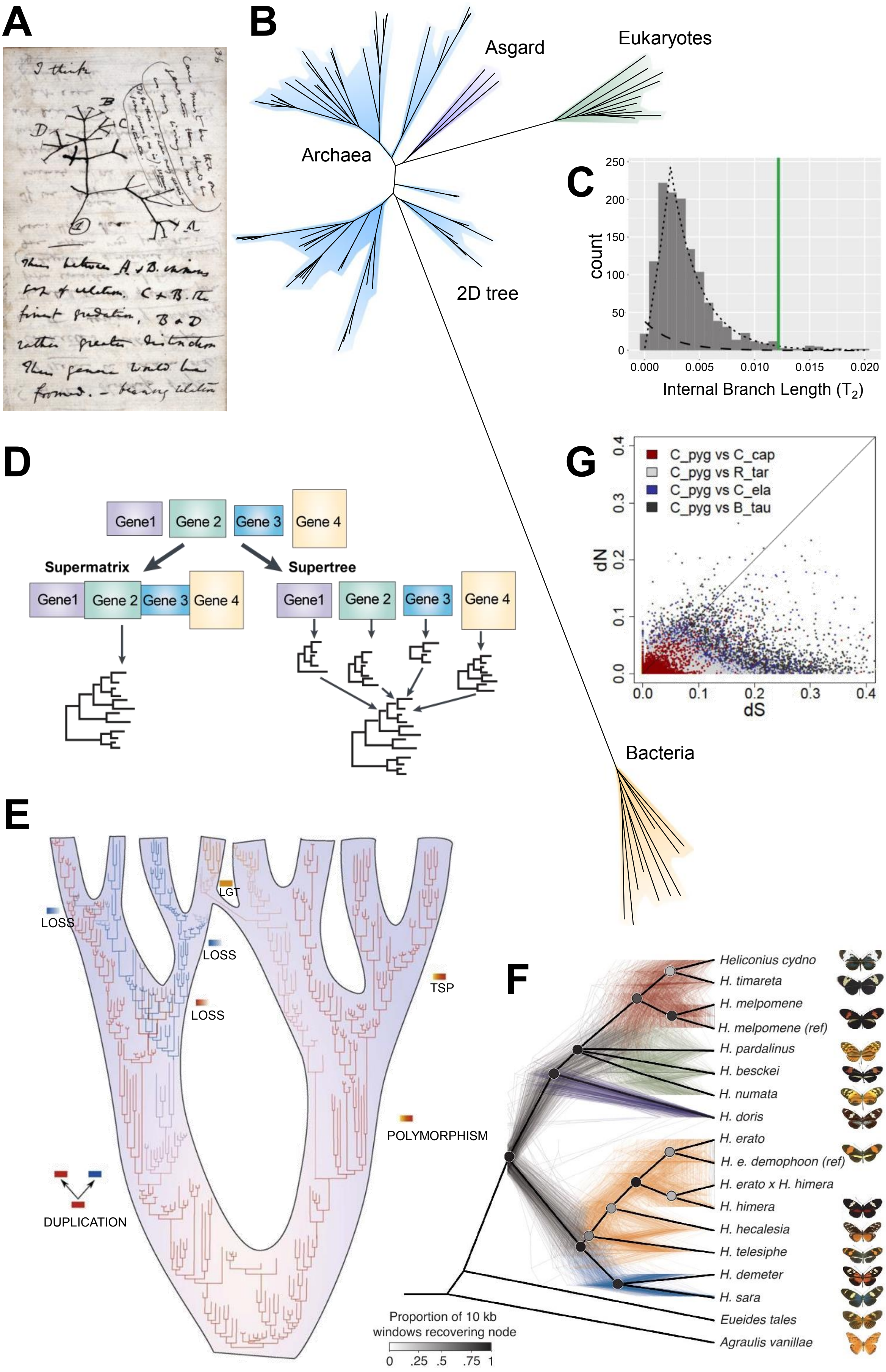
A
I think
B
D
C
A
Thus between A & B immense gap of relation. C & B the finest gradation, B & D rather greater distinction Thus genera would be formed. — bearing relation
B
Asgard
Eukaryotes
Archaea
2D tree
Bacteria
C
count
250
200
150
100
50
0
0.000
0.005
0.010
0.015
0.020
Internal Branch Length (T_2)
D
Gene1
Gene 2
Gene 3
Gene 4
Supermatrix
Supertree
G
C_pyg vs C_cap
C_pyg vs R_tar
C_pyg vs C_ela
C_pyg vs B_tau
dN
dS
0.0
0.1
0.2
0.3
0.4
E
LOSS
LGT
LOSS
LOSS
TSP
POLYMORPHISM
DUPLICATION
F
Heliconius cydno
H. timareta
H. melpomene
H. melpomene (ref)
H. pardalinus
H. besckei
H. numata
H. doris
H. erato
H. e. demophoon (ref)
H. erato x H. himera
H. himera
H. hecalesia
H. telesiphe
H. demeter
H. sara
Eueides tales
Agraulis vanillae
Proportion of 10 kb windows recovering node
0
.25
.5
.75
1

Figure 3 Applications of phylogenomic and evolutionary genomic analyses. **A** Charles Darwin's "Tree-of-Life" sketch from notebook B, 1837, Cambridge University Library. ms.DaR.121:p36 (Atzmon 2015). **B** Phylogenomics provides robust support for a two-domains tree of life, using supertree and coalescent methods to interrogate >3,000 gene families (Williams et al. 2020). **C** Histogram of internal branch lengths in windows with a certain topology of butterflies (Edelman et al. 2019). The ILS-only distribution is shown as a dashed line, the introgression distribution is shown as a dotted line. The average internal branch length in the inversion is shown as a green vertical line. **D** Flowchart depicting the two alternative approaches (concatenated matrix or consensus) that can be used to infer phylogenetic trees (Delsuc et al. 2005). **E** The history of gene trees depicted within the bounds of a species tree (Boussau and Daubin 2010). Processes acting at the genomic level (duplication, loss, gene transfer) as well as at the population level (polymorphism) are shown. **F** Rooted phylogenetic network including the overall inferred tree hypothesis as well as 500 trees constructed from 10 kilo base pair (kbp) long non-overlapping windows (Edelman et al. 2019). **G** Selection analysis based on a scatterplot of dN and dS values depicting pairwise comparisons of 14,512 genes between the Siberian roe deer and 3 other cervid species and cattle (Jong 2020).

the matrix representation using parsimony (MRP) approach (Ragan 1992) and eventually summary methods of unrooted quartet-trees, consistent under the multi-species coalescent model (MSC) (Mirarab et al. 2014).

In the beginning of phylogenetic reconstructions, polytomies, internal nodes connected to more than two sub-trees, were often regarded as "soft" due to a presumed lack of phylogenetic information and the advent of whole genome sequences promised to overcome these polytomies that were often regarded as "non-resolved" (Gee 2003). However, it quickly became clear that conflicting signals were more abundant in genomic data than previously thought (Figure 3E, Boussau and Daubin 2010; Xu and Yang 2016). Different technical and biological reasons can cause sequences to support deviating topologies, like misalignments or misidentification of non-orthologous sequences, but also ancient introgression and incomplete-lineage sorting (ILS) (Xu and Yang 2016). While technical problems can be addressed, introgression and ILS represent major challenges in phylogenetic research that can, when combined, obscure the true species tree (Hibbins and Hahn 2022). ILS, in which two or more lineages fail to coalesce in their most recent ancestral population looking backwards in time, occurs in all lineages and their frequency is well understood from the neutral MSC model (Hibbins and Hahn 2022). Introgression is unpredictable in this instance, although methods to identify introgression in ancient lineages were recently proposed due to the influence of introgression on the distribution of internal branch lengths (Figure 3C, (Edelman et al. 2019).

Eventually, the existence and abundance of these conflicts in phylogenomic data led to the idea that rapid radiations events do not coincide with a bifurcating tree hypothesis and that these events are best represented in a net-like graph depicting the variety of phylogenetic signals found within genomic data (Figure 3F, Bapteste et al. 2013; Edelman et al. 2019).

3.4.2 Selection research

Ever since early molecular techniques like protein sequencing and gel electrophoresis were established in the 1960s, evolutionary geneticists tried to understand how the now observable substitutions shape phenotypic characteristics and hence the evolution of species. In 1968 and 1969, Kimura as well as King and Jukes presented their theory of neutral evolution that stated, contrary to previous expectations, that most substitutions in the genome are either neutral or deleterious towards fitness (Kimura 1968; King and Jukes 1969). Conversely, this implies that the rate of beneficial mutations that would increase the reproductive success of an individual, here called positive selected, is close to zero. Indeed, after the sequencing of the whole genomes of human and chimpanzee, comparisons verified this expectation (CSAC 2005; Eyre-Walker and Keightley 2009), although exceptions with higher rates exist that might be related to effective population size (Galtier 2016). In case of available protein sequence data, substitutions can be differentiated between non-synonymous and synonymous substitutions. Following this, the ratio of both types of substitutions compared to their potential ratios (d_N/d_S or Ka/Ks) is informative for positive selection in genes, because most mutations in expressed sequences are expected to be deleterious and hence under purifying selection (Hughes et al. 2003). If genes are under positive selection, a rate of $d_N/d_S > 1$ is expected, although this rule of thumb is not applicable in every case and might be misleading when generalized over entire gene-sequences (Nielsen and Yang 1998). Over the last century, this approach was refined by using sliding window-based approaches to not overlook areas under positive selection (Wang et al. 2010) and became a daily-used tool in many different studies that aimed to find genes responsible for a certain observation (Figure 3G, Jong 2020; Kumar et al. 2015; Tollis et al. 2019).

3.5 Thesis objectives

In this thesis, I describe the application of whole genome data from various baleen whales in studies regarding their conservation, speciation, evolution, and genetic resources. Whole genome data will facilitate comprehensive statements about their biology that would be difficult if not impossible based on classical monitoring, given their highly migratory nature in a continuous habitat that encompasses the entire globe. Using whole genome short-read sequencing data and conservation genomic analyses, the impact of the industrial whaling period on genomes of e.g. fin whales (*Balaenoptera physalus*) and blue whales (*Balaenoptera musculus*) was assessed and their current viability will be discussed from a genetic point of view (See Publication 1: Wolf et al. 2022 and Manuscript 1: Wolf et al. submitted). Furthermore, using methods of *de novo* genome assembly and evolutionary genomics, a genome of the elusive southern-circumpolar pygmy right whale (*Caperea marginata*) was compiled and respective data was incorporated into a phylogenomic revision of the baleen whale clade as well as into a multi-species, genome-wide scan for positive selected genes to discuss their potential rapid radiation and their convergent evolution towards gigantism and cancer resistance (See Publication 2: Wolf et al. 2023).

In the first publication, genomes of 51 fin whales from Icelandic waters were sequenced and analyzed that represent three separated intervals in time (1989, 2009 and 2018). Using extensive demographic modeling, the impact of whaling on the overall population size was estimated to complement the otherwise poorly understood pre- and post-whaling stock developments impeded by e.g. the lack of historical records and present monitoring difficulties. Furthermore, all aspects of the reciprocal relationship between genetic diversity, inbreeding and genetic load were measured and their implementations for the Icelandic fin whale population were discussed to point out the need to combine these measures in comprehensive analyses rather than interpreting single aspects of this relationship.

The second publication is focused on the *de novo* genome assembly of the otherwise elusive pygmy right whale which was complemented with a phylogenomic reconstruction and selection analyses. By using fragments of a whole-genome alignment that included nearly all extant species of baleen whales, the distribution of genetic conflicts among the here established most parsimonious tree hypothesis was described and the proportions of conflicts that can be traced back to either incomplete lineage sorting or introgression were estimated. Additionally, a pairwise comparison of selection rates between multiple pairs of baleen whales was conducted that were first identified by phylogenetic targeting. Found genes were eventually discussed together with the results of the phylogenomic reconstruction to formulate hypotheses about the evolution of baleen whales, especially gigantic rorquals.

In the last manuscript, the genomes and mitogenomes of 14 North-Pacific blue whales, six Indo-Australian pygmy blue whales (*B. m. brevicauda*) and one Icelandic sei whale (*Balaenoptera borealis*) were sequenced. Together with publicly available data of 11 North-Atlantic blue whale genomes, different aspects of their genetic speciation process were analyzed to infer if the northern hemisphere blue whale subspecies (*B. m. musculus*) should be separated into different conservation management units or even different subspecies. Furthermore, the genome-wide diversity was measured, and traces of inbreeding were identified by the means of runs of homozygosity (ROH) to, again, discuss the impact of the recent depletion on the genotype of these populations.

4. General Discussion

4.1 Main findings

In my studies on whole genome data of baleen whales I identified three main findings that I will discuss in the following three chapters. First, baleen whales have a somewhat normal or even high genetic diversity contrary to initial expectations given their industrial scaled depletion and their physiology, life expectations and slow reproduction rates. Deciphering the molecular and ecological mechanisms that interplay with genetic diversity and their molecular viability will be the main focus in this chapter.

Second, baleen whale populations from different ocean basins appear to be significantly isolated from each other that resulted in or may still result in parapatric speciation. Discussing the resulting under-classification error, its implications for conservation management and how to mitigate this in the age of genomics will be the topic of this chapter.

Third, the late evolution of baleen whales and especially rorquals appeared to have been a rapid radiation that manifests in a hard-polytomy within phylogenomic analyses. Contextualizing this finding in the overall evolution of whales together with historic changes of ocean dynamics and the potential selection towards gigantic body sizes is the aim of this last chapter.

4.2 Anthropogenic impact on whale genomes

4.2.1 Genetic diversity and whaling

The amount of neutral genetic variation within a population is measured in many different ways, with heterozygosity (He) and mean pairwise nucleotide diversity (π) being the most common ones that roughly coincide in sufficiently panmictic populations (Nei and Li 1979), example of coinciding values in Manuscript 1: Wolf et al. submitted). In baleen whales, genome wide He ranges between 0.02% in the gray whale and 0.2% in the blue whale (See Fig 2, Publication 1: Wolf et al. 2022 and Fig 5, Manuscript 1: Wolf et al. submitted), which is surprising given that the blue whale has one of the longest generation times, lowest reproduction rates and experienced one if not the highest hunting pressure during the industrial whaling era (Taylor 2007). Noteworthy, this level of genetic diversity within the blue whale also represents a relatively high value compared to other vertebrates, especially other mammals (Brüniche-Olsen et al. 2018; Palkopoulou et al. 2015).

Following the most basic Wright–Fisher (WF) model of population evolution by excluding e.g. strong drift or a definitive numbers of sites, the degree of genetic variation is directly dependent on only mutation rate (μ) and effective population size (Ne) ($\pi = 4N\mu$ and $He = 4N\mu \,/\, 4N\mu + 1$) (Kimura 1968, 1971; Nei et al. 1975). This implies that a decrease in the effective population size will automatically decrease He and π. Yet, the observable range of neutral genetic diversity always stays between 0.01 and 20% (Charlesworth and Jensen 2022) and hence surprisingly slim compared to what could be possible given the above formula (up to 75%). This phenomenon was later called Lewontin's Paradox based on its first description by Richard C. Lewontin (Lewontin 1974). While there is an ongoing discussion of which factors may contribute to this paradox (See Charlesworth and Jensen 2022), one factor that is always highlighted is that long-term demographic changes are by far more informative for these values compared to short-term changes in Ne (Charlesworth and Jensen 2022; Teixeira and Huber 2021). This factor is directly measurable in species with short life-cycles and seasonal dependent Ne changes such as naturally occurring *Drosophila* populations (Lange et al. 2022). Given that the peak of industrial whaling happened two to three generations ago, measured from the perspective of blue whale generation times (Taylor 2007), it can therefore be assumed that these changes in population size are too recent to have an impact on the neutral genetic diversity of baleen whale genomes. Instead, all demographic models conducted in the three presented studies suggest a generally high ancient effective population size after the presumed radiation of most modern baleen whales, 4-3 million years ago (See chapter 3), followed by a steady population decline. Within the blue whale, these models predict a longer lasting high Ne which may explain the unusual high value compared to even other baleen whales, especially to the arguably similar fin whale (He of ~1% vs. ~2%, See Table 1 and Fig S5, Publication 1: Wolf et al. 2022, Table 1 and Fig 4, Manuscript 1: Wolf et al. submitted).

This relationship between the recent events and genetic diversity is further weakened by the fact that the here expected drastic change in genetic diversity also requires a drastic bottleneck event, more in the range of 2-20 surviving individuals (Nei et al. 1975), a requirement that is unrealistic given the unclear but definitely higher post-whaling stock size estimations (Roman and Palumbi 2003, See also Fig 1, Publication 1: Wolf et al. 2022) and the high dispersal potential of baleen whales.

4.2.2 Molecular viability of baleen whales

Despite this lack of impact on genetic diversity, the question remains whether this level of genetic diversity is informative for the adaptive potential of baleen whales. Especially in times of extensive anthropogenic influences, a sufficient adaptive potential may be essential for the long-term survival of a species. A precondition for a direct relationship between genetic diversity and adaptive potential is that preexisting higher numbers of neutral or deleterious mutations can shift towards beneficial after external changes and that these mutations have a higher rate of fixation and manifestation in the phenotype afterwards. However, the rate of beneficial mutations that become fixed appears rather independent from genetic diversity and ancient population size compared to the immediate effective population size (Teixeira and Huber 2021). Therefore, the current population sizes and other factors such as the properties, rates, and impact of new mutations might be more informative for the adaptive potential (Teixeira and Huber 2021), all of which are hard to measure, especially in baleen whales. This would also explain existing species with substantially lower genetic diversity without obvious effects on their adaptive capability (Morin et al. 2021; Robinson et al. 2016; Westbury et al. 2018).

Conversely, a greater genetic diversity may actually contribute to a higher deleterious potential, as there may be a greater number of pre-existing recessive deleterious mutations that could become prominent in highly depleted populations due to inbreeding (Kyriazis et al. 2021). However, the evaluation of this requires robust predictions of deleterious mutations, the so-called genetic or mutational load, which is challenging and potentially prone to bioinformatic biases and measurement errors (Simons and Sella 2016, see Fig 4, Publication 1: Wolf et al. 2022 for a comparison based on gene models). Moreover, these mutations are subject to the same random selection of alleles during a bottleneck event and genetic purging, facilitated by inbreeding, may even remove these mutations from the gene pool rapidly (Robinson et al. 2018). Thus, it remains unclear if a high genetic diversity can be directly informative for a higher deleterious potential, but it could be in the case of substantial depletion.

These reflections illustrate the complex, intertwined nature of genetic diversity and molecular viability and that the found high genetic diversity in baleen whales is not directly informative for their fitness. Nevertheless, this does not take away the importance of measuring the genetic diversity of a threatened species, because it can be used to make important conservation decisions given the right context from other measures like inbreeding, genetic load and current stock estimates. Inbreeding is probably the most important factor to mention here, because the timing, extant, and impact on the genome can be directly determined from runs of homozygosity (ROH) and their length distributions (Foote et al. 2021). Also, ROH can be found directly after an inbreeding event without being dependent on a sufficiently long enough time period to manifest within the genome (Gurgul et al. 2016). Another important and straightforward method to make assumptions about the viability of a population is to estimate population sizes and stock dynamics, especially for difficult to monitor species like baleen whales. This can be done, for example, based on the side frequency spectrum and the abundance of rare alleles, as conducted in Fig 1 Publication 1: Wolf et al. 2022 and Table 2 Manuscript 1: Wolf et al. submitted (See also Korneliussen et al. 2013; Liu and Fu 2020). Because a bottleneck event will directly affect this spectrum due to its random subsampling, these measures are again suitable to make assumptions directly after depletion. Neutrality tests based on these alleles like e.g. Tajima's D (Korneliussen et al. 2013) can even be informative for population dynamics if there is sufficient reason to exclude selection as the main driver of these measures (Table 2 Manuscript 1: Wolf et al. submitted). Finally, it can be argued that although genetic load is still difficult to estimate, it will most likely become important for conservation genomics in the future due to expectable progress in gene prediction and more robust estimates of mutation effects on the phenotype (Teixeira and Huber 2021).

Eventually, the research on the molecular viability of baleen whales has led me to conclude that the prospects for baleen whales are rather mixed and varying among different species. The Icelandic fin whale population studied in Publication 1: Wolf et al. 2022 has a rather intermediate genetic diversity of roughly 0.1%. However, only short ROH were found, indicating limited influence of inbreeding and demographic estimates suggested a ~80% population decrease during the industrial whaling, which would be not sufficient enough to influence the genetic diversity if population sizes increase back to a pre-depletion level. In this case, it is not expectable that this population will suffer from molecular long-term effects, provided that the recovery and conservation efforts continue. The three blue whale populations studied in Manuscript 1: Wolf et al. submitted, however, showed a rather opposite picture, with higher genetic diversity of roughly 0.2%, but also substantial inbreeding by the means of long ROH. Although the small numbers of individuals that were sequenced per population did not allow for population stock sizes estimations based on the side frequency spectrum, Tajima's D statistics supported a likely population contraction. This substantial inbreeding and the possibility of a higher deleterious potential given their higher genetic diversity, one can expect a higher likelihood for negative molecular long-term effects that require extensive genomic monitoring and substantial conservation efforts. Especially in baleen whales that for logistic reasons cannot be supported by genetic rescue attempts, a strict protection is more necessary than ever. Nevertheless, their highly migratory behavior may mitigate these effects if individuals are allowed to move freely between different ocean basins.

4.3 Isolation and under-classification error

With the advent of molecular sequencing technologies and the broad application of these methods in cetacean research, it became evident that many of these widely distributed species feature rather high levels of genetic isolation between their populations (Archer et al. 2019; Foote et al. 2016; Lah et al. 2016; Leslie and Morin 2018; Onoufriou et al. 2022). Coinciding, a substantial isolation between northern-hemisphere blue whales from North-Atlantic and North-Pacific was identified, despite both being currently regarded as a single subspecies and handled conjointly in conservation management (IWC 2023, See Manuscript 1: Wolf et al. submitted). Prior to these advances in molecular science, isolation and regional forms were considered unlikely in a seemingly continuous habitat (Taylor, Perrin et al. 2017). Their basic biology and ecology impeded closer observations in a vast and inaccessible open sea, especially in the winter breeding season with poor weather conditions (Taylor, Perrin et al. 2017). Moreover, morphological assessments were a logistical challenge for many whale species due to their enormous size and found differences were usually restricted to size differences and slight color changes that were hard to notice from far away (Archer et al. 2019; Branch, Abubaker et al. 2007).

Any kind of regional isolation will result in the accumulation of new genetic variances and a starting divergence towards speciation that eventually results in incompatibility as discussed above. The question of how speciation happened during the quick radiation of whales (See chapter 3) was often raised, but their answer remains open (Árnason et al. 2018; Foote and Morin 2015; Pastene et al. 2007). All commonly accepted modes of speciation were already discussed to apply in whale evolution. Sympatric speciation is considered likely in some populations of killer whales that developed strong social cultures and hence isolation without physical barriers (Foote and Morin 2015). Classical allopatric speciation was discussed from the perspective of marine ecosystem production that could have established non-visible barriers of gene flow (Branch, Stafford et al. 2007; Pastene et al. 2007). Peripatric speciation, speciation based on founding events, may have had influences in strongly isolated ocean basins like the Mediterranean Sea (Geijer et al. 2016). And eventually, parapatric speciation, with active hybrid zones and only partially limited gene flow, is considered to be the main driver in the development of new regional forms, like also discussed for the northern-hemisphere blue whales (Manuscript 1: Wolf et al. submitted, Archer et al. 2019; Lah et al. 2016; Leslie and Morin 2018; Onoufriou et al. 2022). The question, to which extent which mode facilitated speciation in baleen whales might never be fully answered. In the end, it may have been a combination of all, with changing impacts that changed along with the assumed rather recent selection towards more migratory behavior (See chapter 3).

Regardless of which mode is dominant, it can be expected that many regional forms and taxonomic units are missing in our current understanding of whales, including baleen whales (Taylor, Perrin et al. 2017). Especially on the level of subspecies, this under-classification error in cetaceans may be prevalent based on an estimation made by Taylor and colleagues, who gathered hints of potentially overlooked regional forms (Taylor, Perrin et al. 2017). Such an under-classification error may have important implications for conservation given that geo-political decisions often expect the current taxonomic order to closely resemble the true picture in nature (Haig et al. 2006) and that unnamed groups of individuals are less likely to receive protection (Taylor, Perrin et al. 2017). To systematically approach this problem, guidelines were developed to streamline the description of new subspecies (Taylor, Archer et al. 2017) and different molecular measurements were tested for their goodness-of-fit given certain undisputed populations, subspecies and species (Rosel et al. 2017). Although the authors acknowledged that evolution is not uniform enough to broadly apply these recommendations to other groups of organisms, it found recognition within the scientific community by helping to contextualize molecular based differences (Archer et al. 2019; Onoufriou et al. 2022).

Nevertheless, this approach only poorly fits the here presented study of blue whale divergence, mainly due to two reasons. First, the technical part was lacking a modern whole genome perspective. Not only were the proposed thresholds based on mitochondrial marker sequences and thus not directly comparable to whole genome data, they also did not consider modern and comprehensive gene flow analyses (examples in: Jong et al. 2023), which may have important implications for the biological species definition (Arnold 2016). Second, while all here studied blue whale populations showed extensive isolation and only smaller amounts of genetic exchange, their genomes also featured rather low levels of genetic divergence compared to the few available other whole genome studies published on cetacean populations so far (Archer et al. 2019; Onoufriou et al. 2022). This would usually support a classification as “isolated populations” rather than subspecies, but these findings include the currently well recognized and morphologically different subspecies of the pygmy blue whale (*B. m. brevicauda*). An explanation for this could be that these populations simply required less genetic divergence to become isolated, incompatible, and noticeably different, three main arguments when establishing new species and subspecies. A large effective population size during the split together with an already high genetic diversity and a potentially short time since the split may

have further blurred the picture in this instance as already shown in examples of dolphins (Perrin 1975; Taylor, Archer et al. 2017).
Due to these shortcomings of the systematic approach established by Taylor and colleagues, a revision of this systematic approach that not only adds gene flow analyses but also updates the proposed measurements is much needed to fit modern whole genome data.

4.4 Baleen whale evolution

The evolution of baleen whales could explain many of the genomic features described in the here presented three studies. And like a famous essay from Theodosius Dobzhansky stated: "*Nothing in biology makes sense except in the light of evolution*" (Dobzhansky 1973). Since the first ancient whale, the iconic *Basilosaurus* (meaning "king lizard"), was discovered in 1834 and mistakenly named in the fashion of a dinosaur (Uhen 2009), much progress has been made in the fields of paleontology and phylogenetic systematics that gives us an increasingly better idea of how this remarkably transition from land-living artiodactyls to predominantly pelagic marine animals occurred over the last ~50 million years. Nevertheless, both the poor fossil record of cetaceans and our incomplete understanding of past climate and ocean dynamics leave much room for debate, and I have decided to outline the most common hypotheses, with the disclaimer that future findings may change the current interpretation.

4.4.1 Early evolution of baleen whales (Eocene to Pliocene)

Five major events mark the evolution of modern baleen whales that inevitably shaped their genomes. The first event, the transition into mainly aquatic animals began about ~53 – 45 million years ago (Mya), as indicated by the oldest cetacean group known from the fossil record, the *Pakicetidae*, which are considered to have lived an amphibious lifestyle in rivers near the coast of the ancient Tethys Sea (Fordyce 2018). This lifestyle may have allowed a radiation towards coastal waters, which is indicated by more specialized groups that appear shortly thereafter in the fossil record such as *Ambulocetidae*, *Remingtonocetidae* and *Protocetidae*, the latter being considered the most adapted towards marine life as well as containing the ancestors of all modern whales (Uhen 2010). The main drivers that facilitated a selection towards a fully aquatic life are speculated to be e.g. more abundant marine prey, competition or physical stressors during the glacial cycles (Cabrera et al. 2021). Traces of this rapid morphological change can also be found within their genomes as indicated by the loss of many genes associated with e.g. hair, taste, smell, and color vision (Chen et al. 2013; Meredith et al. 2011; Meredith et al. 2013) and by numerous genes with elevated numbers of mutations found in cetacean species specifically (Chikina et al. 2016; Shen et al. 2012).

The second event marks the radiation of fully pelagic crown whales, Neoceti. The first Neoceti fossils date back to ~40 Mya (Buono et al. 2016) and the diversification in the late Eocene (38-35 Mya) is considered the most rapid among cetaceans as the Neoceti include more than 245 fossil and extant genera (Uhen and Pyenson 2007). Apart from a niche radiation, this diversification is also discussed to have been accelerated by various processes such as shifts of oceanic barriers and currents, but also by the general cooling phase that resulted in the beginning of the Antarctic glaciation and enhanced ocean productivity (Cabrera et al. 2021; Marx and Fordyce 2015). This process also led to the split between today's Mysticeti (baleen whales) and Odontoceti (toothed whales), although ancestral Mysticeti most likely possessed teeth and a suction feeding mechanism (Gatesy et al. 2022; Lambert et al. 2017) and only later evolved the typical baleen plates, by presumably re-functioning enlarged palates (Deméré et al. 2008; Gatesy et al. 2022). Genetically, the loss of teeth can be traced back within the genomes of all modern baleen whales as indicated by a substantial molecular erosion in genes related to tooth development (Gatesy et al. 2022). The exact origin of baleen plates is unknown, but a longer period of coexistence between teeth and baleen has been demonstrated in Mysticeti based on parallel occurring fossil records (Hocking et al. 2017; Marx and Fordyce 2015). Eventually, the diversification peaked at around 28 Mya, including many already considerable large species (Fordyce and Marx 2018; Uhen 2010). This peak coincides with the onset of a gradual development of the Antarctic Circumpolar Current (ACC) which may have had important implications for the evolution of baleen whales (Katz et al. 2011).

The third event, beginning at around 23 Mya, describes a slower, stable diversification of baleen whales and a gradual shift to filter feeding, along with the eventual extinction of toothed Mysticeti (Marx and Fordyce 2015). This phase also includes the establishment of the three modern groups of Mysticeti, the Balaenidae (right whales), Cetotheriidae (the pygmy right whale) and Balaenopteridae (rorquals). It is assumed that the full establishment of the ACC created favorable conditions for filter feeding whales and that the numbers of baleen whales flourished in general (Berger 2007; Katz et al. 2011; Marx and Fordyce 2015). However, the exact reason for the extinction of toothed Mysticeti is unknown and may be due to increased competition from Odontoceti with echolocation and the emergence of pinnipeds (Berta 1991; Geisler et al. 2014). This phase of baleen whale diversification also includes the hard-polytomy in the early evolution of rorquals that was identified within Publication 2 (20 –11 Mya, Figure 3, Publication 2: Wolf et al. 2023). During these analyses, evidence was found for a higher degree of introgression between the basal lineages of rorquals. Given that their emergence might not necessarily represent an adaptive radiation because baleen filtering already existed for a longer period of time (Gatesy et al. 2022), it may be that this speciation was relatively slower, involving a longer period of hybridization. Eventually, the diversification of all baleen whales, including rorquals, slowed in this phase, possibly reflecting the effects of ecological niche filling and remained stable up until the Mid-Pliocene (4 – 3 Mya , Freckleton and Harvey 2006; Marx and Fordyce 2015).

4.4.2 Recent baleen whale evolution (Pliocene to LGM)

The fourth event is the relatively recent extinction of smaller baleen whales along with the emergence of modern, giant whale species such as the blue whale and fin whale. This event also led to a rather dramatic decrease in species diversity of baleen whales roundabout, unproportionally affecting smaller species (Slater et al. 2017). The reason for this shift potentially lies in the beginning decline of temperatures during the Pliocene which went on in repeating glacial cycles during the Pleistocene (Schepper et al. 2014). These repeated redirection of ocean currents may have caused repeated shifts of suitable habitats favoring long, often seasonal migrations and larger sizes to overcome larger distances and longer periods without food (Marx and Fordyce 2015). Furthermore, glacial dynamics may have caused a large-scale erosion of iron-rich bedrock, resulting in an increased fertility of the Arctic and Antarctic oceans, allowing for larger body sizes (Martínez-García et al. 2014). The demographic models in all three here presented studies may indicate these favorable conditions for large modern whale species in a relatively high abundance 4 – 2 Mya, followed by a general decline that may reflect the glacial cycles (Fig S5, Publication 1: Wolf et al. 2022, Fig. 4 Publication 2: Wolf et al. 2023, Fig. 4 Manuscript 1: Wolf et al. submitted).

Given that seasonal migrations still exist (Silva et al. 2013) and that these shifts of food sources may still be active with respect to the current climate change, it can be assumed that a general selection trend towards migration and hence larger body sizes still exists. Such a trend towards gigantic body sizes might require adaptations towards cancer resistance since higher amounts of cells and cell divisions would inevitably increase the risk of tumor development, following the assumptions behind Peto’s paradox (Peto et al. 1975; Silva et al. 2023; Tollis et al. 2019). In the second study, a genome wide comparison between multiple pairs of baleen whales with diverging body sizes was conducted in order to find shared patterns of positive selection (Publication 2: Wolf et al. 2023). Although the analyses identified more than 200 genes with significant positive selection, few genes were positively selected in all large whale species

(Table 2, Publication 2: Wolf et al. 2023). Furthermore, previously conducted similar studies rarely to never reported the same genes (Keane et al. 2015; Silva et al. 2023; Tollis et al. 2019). From an evolutionary perspective, namely that the major lineages of modern whales already existed since the baleen whale diversification period (20 – 11 Mya) and that the selection towards larger sizes is comparatively recent (4 – 3 Mya), it can be argued that this lack of congruence is best explained by a convergent evolution towards larger size as already proposed based on fossil records (Slater et al. 2017).

These findings could have important implications for further research on this topic, both in genomics and medical research. The direct translation from genotypic to phenotypic changes is still poorly understood, and finding common patterns is usually a helpful approach to find genes with the greatest impact on a particular phenotypic trait in a complex network of potentially involved genes (Dyson et al. 2022; Shinzato et al. 2021). As a result, genome comparisons across all baleen whales, spanning over multiple origins of gigantism, may miss important genotypic changes in specific lineages. Therefore, it may be more profitable to focus on monophyletic lineages with a single occurrence of gigantism, such as comparing selection rates in the relatively young clade of sei whale (*B. borealis*), bryde whale (*B. brydei, B. edeni, B. ricei*) and the much larger blue whale (*B. musculus*).

4.4.3 Evolutionary dynamics in modern populations (LGM - Anthropocene)

The Last Glacial Maximum (LGM) marks the last natural event that shaped the evolution of baleen whales, their population sizes and genomes, and has since then created relatively stable conditions with an increasing ocean productivity along with increasing temperatures around the Pleistocene-Holocene transition 12 – 7 thousand years ago (kya) (Tsandev et al. 2008). It is noteworthy, in this instance, that while cold water is generally considered to be more nutrient-rich and hence more productive due to ocean dynamics such as upwelling, that a small temperature increase still increases ocean productivity, provided that it does not affect upwelling of nutrients (Tsandev et al. 2008). Accordingly, models based on genetic evidence suggest an increasing population size for all baleen whale species after the LGM (Cabrera et al. 2022).

Although this phase cannot be resolved with the demographic models presented here, it may have had two distinct influences on baleen whale genomes that were visible within the results of the here presented studies. First, this long-term increase in population sizes may have shaped the genetic diversity we see in present-day genomes (Publication 1: Wolf et al. 2022, Manuscript 1: Wolf et al. submitted), as discussed above in Chapter 1 of the Discussion.

Second, the more stable climatic phase may have resulted in a reduced need for migration, allowing populations to be regionally restricted which may facilitate the higher degrees of isolation we see today. Especially for the blue whale it was already discussed that the offset of the LGM also marks the establishment of an isolation population in Indo-Australian waters (Attard et al. 2015), eventually forming the distinct pygmy blue whale (*B. m. brevicauda*) and possibly other northern subspecies as discussed in Manuscript 1: Wolf et al. submitted.

4.5 Conclusion and Outlook

My studies on baleen whale conservation genomics revealed the non-trivial nature of genetic diversity in whale conservation. The time period since industrial whaling is too short and the physiology, life cycle and mutation rates too slow to impact the neutral genetic diversity just yet. However, this does not exclude potential long-term consequences and studies on their molecular viability are more important than ever to ensure their beginning recovery. Instead of focusing on neutral genetic diversity alone, this dissertation outlines a combination of measures that not only change rather quickly after an occurring bottleneck event but that were also more informative in the here studied whale populations. The results also showcased the different impacts of the poorly documented whaling period on different baleen whale species. This calls for more and extensive conservation genomic estimations of all affected whale species and the here provided insights should help to establish a framework to comprehensively study the molecular viability of baleen whales.

The study of baleen whale conservation genomics also revealed that despite their high dispersal potential and the lack of obvious morphological differences, isolated populations and subspecies exist and many may still be unknown to conservation. A framework to comprehensively analyze and define baleen whale subspecies based on molecular data was already attempted but is lacking a modern whole-genome perspective. This includes not only outdated thresholds and potential new ways to measure genetic divergence, but also a complete lack of gene flow analyses that can only be done comprehensively using whole-genome data. Albeit this framework provided a much-needed objective and reproducible way to compare whale populations, a revision of this framework is necessary to keep up with the recent theoretical and technological advances.

Evolutionary genomics allowed the characterization of the hard-polytomy in rorquals that explains the many conflicting assignments of species like the gray whale (*Eschrichtius robustus*), even in past phylogenomic studies. Although the here presented analysis includes nearly all extant taxa and describes the major evolutionary conflict in baleen whales, much remains to be uncovered when focusing on more recent events. Especially the relationships between blue whales, sei whales and whales of the bryde whale complex remain poorly understood and vastly differ depending on the type of data, amount of data and method to analyze the data (McGowen et al. 2020; Rosel et al. 2021). So far, no study included data of all presumed species at once and no study attempted to describe the amount of conflicts at these branches that may be abundant given that even phylgenomic datasets were not able to yield high support values (McGowen et al. 2020). This group of whales is also a promising target for more in depth studies of Peto’s paradox. As pointed out in the here presented study as well as by multiple paleontologists, it is likely that baleen whales contain multiple independent origins of gigantism and hence tumor resistance that impede the usual approach of looking for shared genetic patterns. Instead, it is probably more profitable to look at single origins of gigantism and although the exact identification of these origins remains to be solved in the future, one such origin might be within the clade of blue whale, sei whale and bryde whale complex. Because this clade contains not only potentially multiple independent pairs of species, but also one of the largest size differences that evolved in a considerably short evolutionary time period, it may be the best candidate for further research on this topic.

5. References

Andrews KS, Nichols KM, Elz A, Tolimieri N, Harvey CJ, Pacunski R, Lowry D, Yamanaka KL, Tonnes DM. 2018. Cooperative research sheds light on population structure and listing status of threatened and endangered rockfish species. *Conserv Genet*. 19(4):865–878.

Archer FI, Brownell RL, Hancock-Hanser BL, et al. (9 co-authors). 2019. Revision of fin whale Balaenoptera physalus (Linnaeus, 1758) subspecies using genetics. *Journal of Mammalogy*. 100(5):1653–1670.

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. *Science advances*. 4(4):eaap9873.

Arnold ML. 2016. Divergence with genetic exchange, Oxford, United Kingdom: Oxford University Press.

Attard CRM, Beheregaray LB, Jenner KCS, Gill PC, Jenner M-NM, Morrice MG, Teske PR, Möller LM. 2015. Low genetic diversity in pygmy blue whales is due to climate-induced diversification rather than anthropogenic impacts. *Biology letters*. 11(5):20141037.

Attard CRM, Beheregaray LB, Sandoval-Castillo J, Jenner KCS, Gill PC, Jenner M-NM, Morrice MG, Möller LM. 2018. From conservation genetics to conservation genomics: a genome-wide assessment of blue whales (Balaenoptera musculus) in Australian feeding aggregations. *Royal Society open science*. 5(1):170925.

Atzmon L. 2015. Intelligible design: the origin and visualization of species. *Communication Design*. 3(2):142–156.

Bapteste E, van Iersel L, Janke A, et al. (8 co-authors). 2013. Networks: expanding evolutionary thinking. *Trends in genetics: TIG*. 29(8):439–441.

Berger WH. 2007. Cenozoic cooling, Antarctic nutrient pump, and the evolution of whales. *Deep Sea Research Part II: Topical Studies in Oceanography*. 54(21-22):2399–2421.

Berta A. 1991. New Enaliarctos* (Pinnipedimorpha) from the Oligocene and Miocene of Oregon and the Role of "Enaliarctids" in Pinniped Phylogeny. *Smithsonian Contributions to Paleobiology*. (69):1–33.

Bertorelle G, Raffini F, Bosse M, Bortoluzzi C, Iannucci A, Trucchi E, Morales HE, van Oosterhout C. 2022. Genetic load: genomic estimates and applications in non-model animals. *Nature reviews. Genetics*. 23(8):492–503.

Bininda-Emonds ORP. 2004. The evolution of supertrees. *Trends in ecology & evolution*. 19(6):315–322.

Bleidorn C. 2017. Phylogenomics, Cham: Springer International Publishing.

Booy G, Hendriks RJJ, Smulders MJM, Groenendael JM, Vosman B. 2000. Genetic Diversity and the Survival of Populations. *Plant Biology*. 2(4):379–395.

Boussau B, Daubin V. 2010. Genomes as documents of evolutionary history. *Trends in ecology & evolution*. 25(4):224–232.

Branch T, Abubaker EMN, Mkango S, Butterworth DS. 2007. Separating Southern Blue Whale Subspecies Based On Length Frequencies Of Sexually Mature Females. *Mar Mam Sci*. 23(4):803–833.

Branch T, Stafford K, Palacios, DM, et al. (38 co-authors). 2007. Past and present distribution, densities and movements of blue whales Balaenoptera musculus in the Southern Hemisphere and northern Indian Ocean. *Mammal Review*. 37(2):116–175.

Brüniche-Olsen A, Kellner KF, Anderson CJ, DeWoody JA. 2018. Runs of homozygosity have utility in mammalian conservation and evolutionary studies. *Conserv Genet*. 19(6):1295–1307.

Bukhman YV, Morin PA, Meyer S, et al. (33 co-authors). 2022. A high-quality blue whale genome, segmental duplications, and historical demography.

Buono MR, Fernández MS, Reguero MA, Marenssi SA, Santillana SN, Mörs T. 2016. Eocene Basilosaurid Whales from the La Meseta Formation, Marambio (Seymour) Island, Antarctica. *Ameghiniana*. 53(3):296.

Butti C, Sherwood CC, Hakeem AY, Allman JM, Hof PR. 2009. Total number and volume of Von Economo neurons in the cerebral cortex of cetaceans. *The Journal of comparative neurology*. 515(2):243–259.

Cabrera AA, Bérubé M, Lopes XM, et al. (7 co-authors). 2021. A Genetic Perspective on Cetacean Evolution. *Annu. Rev. Ecol. Evol. Syst.* 52(1):131–151.

Cabrera AA, Schall E, Bérubé M, et al. (34 co-authors). 2022. Strong and lasting impacts of past global warming on baleen whales and their prey. *Global change biology*. 28(8):2657–2677.

Charlesworth B, Jensen JD. 2022. How Can We Resolve Lewontin's Paradox? *Genome biology and evolution*. 14(7).

Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. *Nature reviews. Genetics*. 10(11):783–796.

Chen Z, Wang Z, Xu S, Zhou K, Yang G. 2013. Characterization of hairless (Hr) and FGF5 genes provides insights into the molecular basis of hair loss in cetaceans. *BMC evolutionary biology*. 13:34.

Chikina M, Robinson JD, Clark NL. 2016. Hundreds of Genes Experienced Convergent Shifts in Selective Pressure in Marine Mammals. *Molecular biology and evolution*. 33(9):2182–2192.

Cooke JG. 2018a. Balaenoptera musculus. *IUCN Red List of Threatened Species*. (e.T2477A15692358).

Cooke JG. 2018b. Balaenoptera physalus. *IUCN Red List of Threatened Species*. (e.T2478A50349982.).

CSAC. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature*. 437(7055):69–87.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nature reviews. Genetics*. 6(5):361–375.

Deméré TA, McGowen MR, Berta A, Gatesy J. 2008. Morphological and molecular evidence for a stepwise evolutionary transition from teeth to baleen in mysticete whales. *Systematic biology*. 57(1):15–37.

Dobzhansky T. 1973. Nothing in Biology Makes Sense except in the Light of Evolution. *The American Biology Teacher*. 35(3):125–129.

Double MC, Andrews-Goff V, Jenner KCS, Jenner M-N, Laverick SM, Branch TA, Gales NJ. 2014. Migratory movements of pygmy blue whales (Balaenoptera musculus brevicauda) between Australia and Indonesia as revealed by satellite telemetry. *PloS one*. 9(4):e93578.

Dyson CJ, Pfennig A, Ariano-Sánchez D, Lachance J, Mendelson Iii JR, Goodisman MAD. 2022. Genome of the endangered Guatemalan Beaded Lizard, Heloderma charlesbogerti, reveals evolutionary relationships of squamates and declines in effective population sizes. *G3 (Bethesda, Md.)*. 12(12).

Edelman NB, Frandsen PB, Miyagi M, et al. (26 co-authors). 2019. Genomic architecture and introgression shape a butterfly radiation. *Science (New York, N.Y.)*. 366(6465):594–599.

Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Molecular biology and evolution*. 26(9):2097–2108.

Fagan WF, Holmes EE. 2006. Quantifying the extinction vortex. *Ecology letters*. 9(1):51–60.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*. 17(6):368–376.

Felsenstein J. 1985. Confidence Limits On Phylogenies: An Approach Using The Bootstrap. *Evolution; international journal of organic evolution*. 39(4):783–791.

Foote AD, Hooper R, Alexander A, et al. (31 co-authors). 2021. Runs of homozygosity in killer whale genomes provide a global record of demographic histories. *Molecular ecology*. 30(23):6162–6177.

Foote AD, Morin PA. 2015. Sympatric speciation in killer whales? *Heredity*. 114(6):537–538.

Foote AD, Vijay N, Ávila-Arcos MC, et al. (17 co-authors). 2016. Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature communications*. 7:11693.

Fordyce RE. 2018. Cetacean Evolution. In: Würsig B, Thewissen JGM, Kovacs KM, editors. Encyclopedia of Marine Mammals: Elsevier. p. 180–185.

Fordyce RE, Marx FG. 2018. Gigantism Precedes Filter Feeding in Baleen Whale Evolution. *Current biology : CB*. 28(10):1670-1676.e2.

Frankham R, Ballou JD, Ralls K, Eldridge MDB, Dudash MR, Fenster CB, Lacy RC, Sunnucks P. 2017. Genetic management of fragmented animal and plant populations, Oxford, New York, NY: Oxford University Press.

Freckleton RP, Harvey PH. 2006. Detecting non-Brownian trait evolution in adaptive radiations. *PLoS biology*. 4(11):e373.

Funk WC, Forester BR, Converse SJ, Darst C, Morey S. 2019. Improving conservation policy with genomics: a guide to integrating adaptive potential into U.S. Endangered Species Act decisions for conservation practitioners and geneticists. *Conserv Genet*. 20(1):115–134.

Gabriel W, Lynch M, Bürger R. 1993. Muller's Ratchet and Mutational Meltdowns. *Evolution; international journal of organic evolution*. 47(6):1744–1757.

Galtier N. 2016. Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis. *PLoS genetics*. 12(1):e1005774.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular biology and evolution*. 14(7):685–695.

Gatesy J, Ekdale EG, Deméré TA, Lanzetti A, Randall J, Berta A, El Adli JJ, Springer MS, McGowen MR. 2022. Anatomical, Ontogenetic, and Genomic Homologies Guide Reconstructions of the Teeth-to-Baleen Transition in Mysticete Whales. *J Mammal Evol*. 29(4):891–930.

Gee H. 2003. Evolution: ending incongruence. *Nature*. 425(6960):782.

Geijer CK, Di Notarbartolo Sciara G, Panigada S. 2016. Mysticete migration revisited: are Mediterranean fin whales an anomaly? *Mam Rev*. 46(4):284–296.

Geisler JH, Colbert MW, Carew JL. 2014. A new fossil species supports an early origin for toothed whale echolocation. *Nature*. 508(7496):383–386.

Goffeau A, Barrell BG, Bussey H, et al. (13 co-authors). 1996. Life with 6000 genes. *Science (New York, N.Y.)*. 274(5287):546, 563-7.

Gordon AD. 1986. Consensus Supertrees. The Synthesis of Rooted Trees Containing Overlapping Sets of Labeles Leaves. *Journal of Classification*. 3:335–348.

Gurgul A, Szmatoła T, Topolski P, Jasielczuk I, Żukowski K, Bugno-Poniewierska M. 2016. The use of runs of homozygosity for estimation of recent inbreeding in Holstein cattle. *Journal of applied genetics*. 57(4):527–530.

Haig SM, Beever EA, Chambers SM, et al. (10 co-authors). 2006. Taxonomic Considerations in Listing Subspecies Under the U.S. Endangered Species Act. *Conservation Biology*. 20(6):1584–1594.

Hennig W. 1965. Phylogenetic Systematics. *Annu. Rev. Entomol.* 10(1):97–116.

Hibbins MS, Hahn MW. 2022. Phylogenomic approaches to detecting and characterizing introgression. *Genetics*. 220(2).

Hocking DP, Marx FG, Fitzgerald EMG, Evans AR. 2017. Ancient whales did not filter feed with their teeth. *Biology letters*. 13(8).

Hucke-Gaete R, Bedriñana-Romano L, Viddi FA, Ruiz JE, Torres-Florez JP, Zerbini AN. 2018. From Chilean Patagonia to Galapagos, Ecuador: novel insights on blue whale migratory pathways along the Eastern South Pacific. *PeerJ*. 6:e4695.

Hughes AL, Packer B, Welch R, Bergen AW, Chanock SJ, Yeager M. 2003. Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proceedings of the National Academy of Sciences of the United States of America*. 100(26):15754–15757.

IWC. 1982. Chairs Report, 34th Annual Meeting --- Brighton. *International Whaling Commision*.

IWC. 2023. Blue whale. Balaenoptera musculus, https://iwc.int/about-whales/whale-species/blue-whale.

Janke A, Feldmaier-Fuchs G, Thomas WK, Haeseler A von, Pääbo S. 1994. The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics*. 137(1):243–256.

Jong MJ de. 2020. Detection of genomic signatures of selection in roe deer and reindeer populations., Durham, UK.

Jong MJ de, Niamir A, Wolf M, et al. (7 co-authors). 2023. Range-wide whole-genome resequencing of the brown bear reveals drivers of intraspecies divergence. *Communications biology*. 6(1):153.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In: Munro HN, Allison JB, editors. Mammalian protein metabolism. New York: Acad. Pr.

Katz ME, Cramer BS, Toggweiler JR, Esmay G, Liu C, Miller KG, Rosenthal Y, Wade BS, Wright JD. 2011. Impact of Antarctic Circumpolar Current development on late Paleogene ocean structure. *Science (New York, N.Y.)*. 332(6033):1076–1079.

Keane M, Semeiks J, Webb AE, et al. (27 co-authors). 2015. Insights into the evolution of longevity from the bowhead whale genome. *Cell reports*. 10(1):112–122.

Kemper CM. 2009. Pygmy Right Whale. In: Perrin W.F., Würsig B.G., Thewissen J.G.M., editors. Encyclopedia of Marine Mammals: Elsevier. p. 939–941.

Kim BY, Huber CD, Lohmueller KE. 2018. Deleterious variation shapes the genomic landscape of introgression. *PLoS genetics*. 14(10):e1007741.

Kimura M. 1968. Evolutionary Rate at the Molecular Level. *Nature*. 217:624–626.

Kimura M. 1971. Theoretical foundation of population genetics at the molecular level. *Theoretical population biology*. 2(2):174–208.

King JL, Jukes TH. 1969. Non-Darwinian evolution. *Science (New York, N.Y.)*. 164(3881):788–798.

Knoop V, Müller K. 2009. Gene und Stammbäume. Ein Handbuch zur molekularen Phylogenetik, Heidelberg: Spektrum Akademischer Verlag.

Korneliussen TS, Moltke I, Albrechtsen A, Nielsen R. 2013. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC bioinformatics*. 14:289.

Kozakiewicz CP, Burridge CP, Funk WC, et al. (12 co-authors). 2019. Urbanization reduces genetic connectivity in bobcats (Lynx rufus) at both intra- and interpopulation spatial scales. *Molecular ecology*. 28(23):5068–5085.

Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic biology*. 56(1):17–24.

Kumar V, Kutschera VE, Nilsson MA, Janke A. 2015. Genetic signatures of adaptation revealed from transcriptome sequencing of Arctic and red foxes. *BMC genomics*. 16(1):585.

Kyriazis CC, Wayne RK, Lohmueller KE. 2021. Strongly deleterious mutations are a primary determinant of extinction risk due to inbreeding depression. *Evolution letters*. 5(1):33–47.

Lah L, Trense D, Benke H, et al. (11 co-authors). 2016. Spatially Explicit Analysis of Genome-Wide SNPs Detects Subtle Population Structure in a Mobile Marine Mammal, the Harbor Porpoise. *PloS one*. 11(10):e0162792.

Lambert O, Martínez-Cáceres M, Bianucci G, Di Celma C, Salas-Gismondi R, Steurbaut E, Urbina M, Muizon C de. 2017. Earliest Mysticete from the Late Eocene of Peru Sheds New Light on the Origin of Baleen Whales. *Current biology : CB*. 27(10):1535-1541.e2.

Lammers F, Blumer M, Rücklé C, Nilsson MA. 2019. Retrophylogenomics in rorquals indicate large ancestral population sizes and a rapid radiation. *Mobile DNA*. 10:5.

Lange JD, Bastide H, Lack JB, Pool JE. 2022. A Population Genomic Assessment of Three Decades of Evolution in a Natural Drosophila Population. *Molecular biology and evolution*. 39(2).

Leslie MS, Morin PA. 2018. Structure and phylogeography of two tropical predators, spinner (Stenella longirostris) and pantropical spotted (S. attenuata) dolphins, from SNP data. *Royal Society open science*. 5(4):171615.

Lewontin RC. 1974. The genetic basis of evolutionary change., New York: Columbia University Press.

Linard B, Ebersberger I, McGlynn SE, et al. (10 co-authors). 2021. Ten Years of Collaborative Progress in the Quest for Orthologs. *Molecular biology and evolution*. 38(8):3033–3045.

Liu X, Fu Y-X. 2020. Stairway Plot 2: demographic history inference with folded SNP frequency spectra. *Genome biology*. 21(1):280.

Martínez-García A, Sigman DM, Ren H, Anderson RF, Straub M, Hodell DA, Jaccard SL, Eglinton TI, Haug GH. 2014. Iron fertilization of the Subantarctic ocean during the last ice age. *Science (New York, N.Y.)*. 343(6177):1347–1350.

Marx FG, Fordyce RE. 2015. Baleen boom and bust: a synthesis of mysticete phylogeny, diversity and disparity. *Royal Society open science*. 2(4):140434.

Masel J. 2011. Genetic drift. *Current biology : CB*. 21(20):R837-8.

McClain CR, Balk MA, Benfield MC, et al. (13 co-authors). 2015. Sizing ocean giants: patterns of intraspecific size variation in marine megafauna. *PeerJ*. 3:e715.

McDonald MA, Mesnick SL, Hildebrand JA. 2006. Biogeographic characterization of blue whale song worldwide. Using song to identify populations. *Journal of Cetacean Research and Management*. 8(1):55–65.

McGowen MR, Tsagkogeorga G, Álvarez-Carretero S, et al. (8 co-authors). 2020. Phylogenomic Resolution of the Cetacean Tree of Life Using Target Sequence Capture. *Systematic biology*. 69(3):479–501.

Meredith RW, Gatesy J, Cheng J, Springer MS. 2011. Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales. *Proceedings. Biological sciences*. 278(1708):993–1002.

Meredith RW, Gatesy J, Emerling CA, York VM, Springer MS. 2013. Rod monochromacy and the coevolution of cetacean retinal opsins. *PLoS genetics*. 9(4):e1003432.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics (Oxford, England)*. 30(17):i541-8.

Morin PA, Archer FI, Avila CD, et al. (31 co-authors). 2021. Reference genome and demographic history of the most endangered marine mammal, the vaquita. *Molecular ecology resources*. 21(4):1008–1020.

Nakamura T, Yamada KD, Tomii K, Katoh K. 2018. Parallelization of MAFFT for large-scale multiple sequence alignments. *Bioinformatics (Oxford, England)*. 34(14):2490–2492.

Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*. 48(3):443–453.

Nei M, Li WH. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*. 76(10):5269–5273.

Nei M, Maruyama T, Chakraborty R. 1975. The Bottleneck Effect and Genetic Variability in Populations. *Evolution*. 29(1):1.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics*. 148(3):929–936.

Onoufriou AB, Gaggiotti OE, Aguilar de Soto N, et al. (38 co-authors). 2022. Biogeography in the deep: Hierarchical population genomic structure of two beaked whale species. *Global Ecology and Conservation*. 40:e02308.

Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW. 2010. Conservation genetics in transition to conservation genomics. *Trends in genetics: TIG*. 26(4):177–187.

Palkopoulou E, Mallick S, Skoglund P, et al. (9 co-authors). 2015. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current biology: CB*. 25(10):1395–1400.

Palumbi SR. 1994. Genetic Divergence, Reproductive Isolation, And Marine Speciation. *Annu. Rev. Ecol. Syst.* 25(1):547–572.

Pareek CS, Smoczynski R, Tretyn A. 2011. Sequencing technologies and genome sequencing. *Journal of applied genetics*. 52(4):413–435.

Pastene LA, Goto M, Kanda N, et al. (8 co-authors). 2007. Radiation and speciation of pelagic organisms during periods of global warming: the case of the common minke whale, Balaenoptera acutorostrata. *Molecular ecology*. 16(7):1481–1495.

Perrin WF. 1975. Variation of spotted and spinner porpoise in the Eastern Pacific and Hawaii. *Bulletin of the Scripps Institution of Oceanography, University of California, Sen Diego La Jolla, California*. 21.

Petit RJ, Excoffier L. 2009. Gene flow and species delimitation. *Trends in ecology & evolution*. 24(7):386–393.

Peto R, Roe FJ, Lee PN, Levy L, Clack J. 1975. Cancer and ageing in mice and men. *British journal of cancer*. 32(4):411–426.

Prost S, Winter S, Raad J de, et al. (8 co-authors). 2020. Education in the genomics era: Generating high-quality genome assemblies in university courses. *GigaScience*. 9(6).

Ragan MA. 1992. Phylogenetic inference based on matrix representation of trees. *Molecular phylogenetics and evolution*. 1(1):53–58.

Ratnarajah L, Melbourne-Thomas J, Marzloff MP, Lannuzel D, Meiners KM, Chever F, Nicol S, Bowie AR. 2016. A preliminary model of iron fertilisation by baleen whales and Antarctic krill in the Southern Ocean: Sensitivity of primary productivity estimates to parameter uncertainty. *Ecological Modelling*. 320:203–212.

Rekdahl ML, Garland EC, Carvajal GA, King CD, Collins T, Razafindrakoto Y, Rosenbaum H. 2018. Culturally transmitted song exchange between humpback whales (Megaptera novaeangliae) in the southeast Atlantic and southwest Indian Ocean basins. *Royal Society open science*. 5(11):172305.

Robinson JA, Brown C, Kim BY, Lohmueller KE, Wayne RK. 2018. Purging of Strongly Deleterious Mutations Explains Long-Term Persistence and Absence of Inbreeding Depression in Island Foxes. *Current biology : CB*. 28(21):3487-3494.e4.

Robinson JA, Ortega-Del Vecchyo D, Fan Z, Kim BY, vonHoldt BM, Marsden CD, Lohmueller KE, Wayne RK. 2016. Genomic Flatlining in the Endangered Island Fox. *Current biology : CB*. 26(9):1183–1189.

Roman J, Palumbi SR. 2003. Whales before whaling in the North Atlantic. *Science (New York, N.Y.)*. 301(5632):508–510.

Ronaghi M. 2001. Pyrosequencing Sheds Light on DNA Sequencing. *Genome Res*. 11(1):3–11.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)*. 19(12):1572–1574.

Rosel PE, Hancock-Hanser BL, Archer FI, et al. (8 co-authors). 2017. Examining metrics and magnitudes of molecular genetic differentiation used to delimit cetacean subspecies based on mitochondrial DNA control region sequences. *Mar Mam Sci*. 33(S1):76–100.

Rosel PE, Wilcox LA, Yamada TK, Mullin KD. 2021. A new species of baleen whale (Balaenoptera) from the Gulf of Mexico, with a review of its geographic distribution. *Marine Mammal Science*. 37(2):577–610.

Sanger F, Coulson AR. 1975. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology*. 94(3):441–448.

Savoca MS, Czapanskiy MF, Kahane-Rapport SR, et al. (14 co-authors). 2021. Baleen whale prey consumption based on high-resolution foraging measurements. *Nature*. 599(7883):85–90.

Schepper S de, Gibbard PL, Salzmann U, Ehlers J. 2014. A global synthesis of the marine and terrestrial evidence for glaciation during the Pliocene Epoch. *Earth-Science Reviews*. 135:83–102.

Schneider C, Woehle C, Greve C, D'Haese CA, Wolf M, Hiller M, Janke A, Bálint M, Huettel B. 2021. Two high-quality de novo genomes from single ethanol-preserved specimens of tiny metazoans (Collembola). *GigaScience*. 10(5).

Seth J von, Dussex N, Díez-Del-Molino D, et al. (18 co-authors). 2021. Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nature communications*. 12(1):2393.

Shapiro B, Rambaut A, Drummond AJ. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular biology and evolution*. 23(1):7–9.

Shen T, Xu S, Wang X, Yu W, Zhou K, Yang G. 2012. Adaptive evolution and functional constraint at TLR4 during the secondary aquatic adaptation and diversification of cetaceans. *BMC evolutionary biology*. 12:39.

Shinzato C, Khalturin K, Inoue J, et al. (7 co-authors). 2021. Eighteen Coral Genomes Reveal the Evolutionary Origin of Acropora Strategies to Accommodate Environmental Changes. *Molecular biology and evolution*. 38(1):16–30.

Silva FA, Souza ÉMS, Ramos E, Freitas L, Nery MF. 2023. The molecular evolution of genes previously associated with large sizes reveals possible pathways to cetacean gigantism. *Scientific reports*. 13(1):67.

Silva MA, Prieto R, Jonsen I, Baumgartner MF, Santos RS. 2013. North Atlantic blue and fin whales suspend their spring migration to forage in middle latitudes: building up energy reserves for the journey? *PloS one*. 8(10):e76507.

Simons YB, Sella G. 2016. The impact of recent population history on the deleterious mutation load in humans and close evolutionary relatives. *Current opinion in genetics & development*. 41:150–158.

Slarkin M. 1985. Gene Flow in Natural Populations. *Annu. Rev. Ecol. Syst.* 16(1):393–430.

Slater GJ, Goldbogen JA, Pyenson ND. 2017. Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Proceedings. Biological sciences*. 284(1855).

Smith CR, Glover AG, Treude T, Higgs ND, Amon DJ. 2015. Whale-fall ecosystems: recent insights into ecology, paleoecology, and evolution. *Annual review of marine science*. 7:571–596.

Spielman D, Brook BW, Frankham R. 2004. Most species are not driven to extinction before genetic factors impact them. *Proceedings of the National Academy of Sciences of the United States of America*. 101(42):15261–15264.

Szewciw LJ, Kerckhove DG de, Grime GW, Fudge DS. 2010. Calcification provides mechanical reinforcement to whale baleen alpha-keratin. Proceedings. Biological sciences. 277(1694):2597–2605.

Taylor BL. 2007. Generation length and precent mature estimates for IUCN assessments of cetaceans. *Southwest Fisheries Science Center Administrative Report*. LJ-07-01.

Taylor BL, Archer FI, Martien KK, et al. (16 co-authors). 2017. Guidelines and quantitative standards to improve consistency in cetacean subspecies and species delimitation relying on molecular genetic data. *Mar Mam Sci*. 33(S1):132–155.

Taylor BL, Perrin WF, Reeves RR, Rosel PE, Wang JY, Cipriano F, Scott Baker C, Brownell RL. 2017. Why we should develop guidelines and quantitative standards for using genetic data to delimit subspecies for data-poor organisms like cetaceans. *Mar Mam Sci*. 33(S1):12–26.

Teixeira JC, Huber CD. 2021. The inflated significance of neutral genetic diversity in conservation genetics. *Proceedings of the National Academy of Sciences of the United States of America*. 118(10).

Tollis M, Robbins J, Webb AE, et al. (9 co-authors). 2019. Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). *Molecular biology and evolution*. 36(8):1746–1763.

Tønnessen JN, Johnsen AO. 1982. The history of modern whaling, Berkeley: University of California Press.

Tsandev I, Slomp CP, van Cappellen P. 2008. Glacial-interglacial variations in marine phosphorus cycling: Implications for ocean productivity. *Global Biogeochem. Cycles*. 22(4):n/a-n/a.

Uhen MD. 2009. Basilosaurids. In: Perrin W.F., Würsig B.G., Thewissen J.G.M., editors. Encyclopedia of Marine Mammals: Elsevier. p. 91–94.

Uhen MD. 2010. The Origin(s) of Whales. *Annu. Rev. Earth Planet. Sci.* 38(1):189–219.

Uhen MD, Pyenson ND. 2007. Diversity Estimates, Biases, and Historiographic Effects. Resolving Cetacean Diversity in the Tertiary. *Palaeontologia Electronica*. 10(2):11A:22p.

van der Valk T, Díez-Del-Molino D, Marques-Bonet T, Guschanski K, Dalén L. 2019. Historical Genomes Reveal the Genomic Consequences of Recent Population Decline in Eastern Gorillas. *Current biology : CB*. 29(1):165-170.e6.

van der Valk T, Manuel M de, Marques-Bonet T, Guschanski K. 2019. Estimates of genetic load suggest frequent purging of deleterious alleles in small populations.

Venter JC, Adams MD, Myers EW, et al. (271 co-authors). 2001. The sequence of the human genome. *Science (New York, N.Y.)*. 291(5507):1304–1351.

Walters AD, Schwartz MK. 2021. Population Genomics for the Management of Wild Vertebrate Populations. In: Hohenlohe PA, Rajora OP, editors. Population Genomics: Wildlife. Cham: Springer International Publishing. p. 419–436.

Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, proteomics & bioinformatics*. 8(1):77–80.

Waples RS, Gaggiotti O. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular ecology*. 15(6):1419–1439.

Waterston RH, Lindblad-Toh K, Birney E, et al. (219 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*. 420(6915):520–562.

Westbury MV, Hartmann S, Barlow A, et al. (8 co-authors). 2018. Extended and Continuous Decline in Effective Population Size Results in Low Genomic Diversity in the World's Rarest Hyena Species, the Brown Hyena. *Molecular biology and evolution*. 35(5):1225–1237.

Wetterstrand KA. 2021. DNA Sequencing Costs. Data from the NHGRI Genome Sequencing Program (GSP), www.genome.gov/sequencingcostsdata.

Wicke S, Schäferhoff B, dePamphilis CW, Müller KF. 2014. Disproportional plastome-wide increase of substitution rates and relaxed purifying selection in genes of carnivorous Lentibulariaceae. *Molecular biology and evolution*. 31(3):529–545.

Williams TA, Cox CJ, Foster PG, Szöllősi GJ, Embley TM. 2020. Phylogenomics provides robust support for a two-domains tree of life. *Nature ecology & evolution*. 4(1):138–147.

Winter S, Prost S, Raad J de, et al. (19 co-authors). 2020. Chromosome-level genome assembly of a benthic associated Syngnathiformes species: the common dragonet, Callionymus lyra. *GigaByte (Hong Kong, China)*. 2020:gigabyte6.

Winter S, Raad J de, Wolf M, et al. (15 co-authors). 2023. A chromosome-scale reference genome assembly of the great sand eel, Hyperoplus lanceolatus. *The Journal of heredity*. 114(2):189–194.

Wolf M, Jong M de, Halldórsson SD, Árnason Ú, Janke A. 2022. Genomic Impact of Whaling in North Atlantic Fin Whales. *Molecular biology and evolution*. 39(5).

Wolf M, Zapf K, Gupta DK, Hiller M, Árnason Ú, Janke A. 2023. The genome of the pygmy right whale illuminates the evolution of rorquals. *BMC biology*. 21(1):79.

Wray J, Keen E, O'Mahony ÉN. 2021. Social survival: Humpback whales (Megaptera novaeangliae) use social structure to partition ecological niches within proposed critical habitat. *PloS one*. 16(6):e0245409.

Xu B, Yang Z. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. *Genetics*. 204(4):1353–1368.

Yadav SP. 2007. The wholeness in suffix -omics, -omes, and the word om. *Journal of biomolecular techniques : JBT*. 18(5):277.

6. Publications

6.1 Publication 1:

Genomic impact of whaling in North Atlantic fin whales.

Magnus Wolf, Menno J. de Jong, Sverrir Daníel Halldórsson, Úlfur Árnason, Axel Janke

Published as a research article in: *Molecular Biology and Evolution*, Volume 39, Issue 5, May 2022, msac094, https://doi.org/10.1093/molbev/msac094

Declaration of author contributions to the publication

Publication: Genomic impact of whaling in North Atlantic fin whales.

Status: published

Name of the journal: Molecular Biology and Evolution

Contributing Authors:

Magnus Wolf (MW), Menno J. de Jong (MdJ)), Sverrir Daníel Halldórsson (SDH), Úlfur Árnason (UA), Axel Janke (AJ)

Contributions of the doctoral candidate and his co-authors:

1.) Concept and design		
MW	65%	
MdJ	10%	
UA	5%	
AJ	20%	
2.) Data analysis and figure compilation		
MW	100%	Compilation of all datasets and figures
3.) Interpretation of data		
MW	80%	Interpretation of all data
AJ	20%	Summary interpretation
4.) Drafting manuscript		
MW	65%	
MdJ	10%	
UA	5%	
AJ	20%	
5.) Other works		
SDH	100%	Collection of samples and recherche of origins

I hereby certify that the information above is correct.

Date and place

Signature doctoral candidate

Date and place

Signature supervisor

Genomic Impact of Whaling in North Atlantic Fin Whales

Magnus Wolf ,*[1,2] Menno de Jong ,[1] Sverrir Daníel Halldórsson ,[3] Úlfur Árnason ,[4,5] and Axel Janke [1,2,6]

[1]Senckenberg Biodiversity and Climate Research Centre (BiK-F), Georg-Voigt-Strasse 14-16, Frankfurt am Main, Germany
[2]Institute for Ecology, Evolution and Diversity, Goethe University, Max-von-Laue-Strasse. 9, Frankfurt am Main, Germany
[3]Marine and Freshwater Research Institute, Fornubúðir 5, IS 220 Hafnarfjörður, Iceland
[4]Department of Clinical Sciences Lund, Lund University, Sweden
[5]Department of Neurosurgery, Skane University Hospital in Lund, Sweden
[6]LOEWE-Centre for Translational Biodiversity Genomics (TBG), Senckenberg Nature Research Society, Georg-Voigt-Straße 14-16, Frankfurt am Main, Germany

***Corresponding author:** E-mail: magnus.wolf@senckenberg.de.

Associate editor: Rebekah Rogers

Abstract

It is generally recognized that large-scale whaling in the 19th and 20th century led to a substantial reduction of the size of many cetacean populations, particularly those of the baleen whales (*Mysticeti*). The impact of these operations on genomic diversity of one of the most hunted whales, the fin whale (*Balaenoptera physalus*), has remained largely unaddressed because of the paucity of adequate samples and the limitation of applicable techniques. Here, we have examined the effect of whaling on the North Atlantic fin whale based on genomes of 51 individuals from Icelandic waters, representing three temporally separated intervals, 1989, 2009 and 2018 and provide a reference genome for the species. Demographic models suggest a noticeable drop of the effective population size of the North Atlantic fin whale around a century ago. The present results suggest that the genome-wide heterozygosity is not markedly reduced and has remained comparable with other baleen whale species. Similarly, there are no signs of apparent inbreeding, as measured by the proportion of long runs of homozygosity, or of a distinctively increased mutational load, as measured by the amount of putative deleterious mutations. Compared with other baleen whales, the North Atlantic fin whale appears to be less affected by anthropogenic influences than other whales such as the North Atlantic right whale, consistent with the presence of long runs of homozygosity and higher levels of mutational load in an otherwise more heterozygous genome. Thus, genome-wide assessments of other species and populations are essential for future, more specific, conservation efforts.

Key words: **fin whales, bottleneck, genetic diversity, runs of homozygosity, whaling, mutational load, demography.**

Introduction

Fin Whales and Whaling

Fin whales (*Balaenoptera physalus*) are a species of cosmopolitan rorquals (*Balaenopteridae*) within the group of baleen whales (*Mysticeti*) (Edwards et al. 2015; Aguilar and García-Vernet 2017). They are among the largest species on Earth and are known for migrating seasonally between low latitude breeding and high latitude feeding grounds (Silva et al. 2013; Lydersen et al. 2020). Despite their global occurrence, fin whales rarely cross the equatorial regions, and their distribution is therefore defined by the equator and major landmasses (Edwards et al. 2015). As these restrictions have existed for long periods of time, they have made it possible to differentiate fin whales into distinct populations and subspecies based on both phenotypic and genotypic features (Lockyer and Waters 1986; Edwards et al. 2015; Archer et al. 2019).

Fin whales have been subjected to large-scaled whaling since first industrialized operations in the 1870s (Aguilar and García-Vernet 2017). In 1904, a first local over-exploitation was reached in the waters of northern Finnmark (Norway) after which the local whaling industry was forced to either switch targets to, for example, the minke whale, *Balaenoptera acutorostrata* or move to other locations (Tønnessen and Johnsen 1982). From there on, whaling expanded around the globe with reoccurring over-exploitations and relocations of industry infrastructure until catch rates peaked between 1925 and 1960 with records of ~30,000 individuals taken annually (Aguilar and García-Vernet 2017). In the 1960s, 1970s and 1980s, increasingly strict whaling limitations were introduced and later a complete moratorium was enforced by the International Whaling Commission due to dwindling stock sizes and imminent extinctions of several baleen whale species (Smith 1984). This led to a noticeable recovery of fin whale population sizes and current estimates add up to ∼90,000–100,000 individuals worldwide of which 40,000–60,000 are allocated to the North Atlantic (Aguilar and García-Vernet 2017; Hansen et al. 2019). A recent survey conducted in the North Atlantic area around Icelandic waters sized up 30,000 individuals for this area alone (Pike et al. 2019). Eventually, the recent recovery led to a change in the threat status of the IUCN red list from "endangered" to "vulnerable" (Cooke 2018).

Genetic Diversity and Conservation

Population survivability is not dependent on census sizes only, but is also shaped by genetic diversity, which is a proxy of the adaptive potential and hence long-term survival of a species (Booy et al. 2000), a circumstance which does not necessarily coincide with abundance (Coates et al. 2018). There are numerous examples of species with relatively high present-day census sizes but low genetic diversity such as the Madagascar fish-eagle, the brown hyena or the narwal (Johnson et al. 2009; Westbury et al. 2018, 2019). Despite these examples, it is commonly assumed that populations with low genetic diversity are more vulnerable to extinction than others because in cases of rapidly changing environments, a low genetic diversity might result in a decreased adaptive potential and hence a lowered reproduction rate and increased mortality (Reed and Frankham 2003; Spielman et al. 2004a; Frankham 2005).

In cases of low genetic diversity and effective population size (N_e), inbreeding may cause an accumulation of homozygous recessive mutations that eventually affect the fitness of a species due to their deleterious effects (Tanaka 2000). The abundance of deleterious mutations, the so-called mutational load, might further increase, because of a reduced efficiency of purifying selection (Ohta 1973). This reciprocal relationship between genetic diversity, inbreeding, mutational load and fitness is widely known as the "extinction vortex" and has been studied in detail, both on theoretical grounds and to guide conservation practices (Charlesworth and Willis 2009; Kimura et al. 1963; Leimu et al. 2006; Spielman et al. 2004b; Tanaka 2000).

Genome-wide studies addressing these topics are still rare, despite their promise to yield comprehensive conclusions on the genetic diversity and the general fitness of a population (Bortoluzzi et al. 2020; von Seth et al. 2021; van der Valk et al. 2019b; Westbury et al. 2018). Moreover, inclusive studies addressing inbreeding or mutational load require well assembled and annotated genomic data or genome information from numerous individuals. Until recently, computational and economic limitations have hindered a common application of whole-genome-sequencing data in conservation biology. However, since sequencing costs are steadily decreasing and computational power is increasing, these limitations are disappearing, enabling a broad and large-scale application of conservation-genomic analyses.

Objectives

In this study, we assess the genomic consequences of industrial whaling for a North Atlantic fin whale population over a time period of three decades, spanning approximately one full generation time of the species. We sequence the genomes of 51 fin whale individuals which were sampled around Iceland in 1989, 2009, and 2018 and provide a new high-quality reference genome assembly for the species. Genome data is used to model the demographic past of the population and to quantify genome-wide heterozygosity as a measure for genetic diversity. To analyze potential genetic consequences, inbreeding factors are calculated based on the distribution of runs of homozygosity (ROH) and the mutational load is estimated by identifying the abundance of potential deleterious mutations. These results are compared to a broad selection of other baleen whale species that experienced different magnitudes of whaling (Tønnessen and Johnsen 1982). With this study we aim to present an overview of the genetic variability in North Atlantic fin whales and demonstrate the need for comprehensive, genome-wide data to assess the genetic impact and consequences of a bottleneck caused by extensive hunting.

Results

Genome Characteristics and Completeness

A high-quality reference genome for the fin whale (*Balaenoptera physalus*) was assembled to a total length of 2.412 Gbp with a contig N50 of 24.9 Mbp and a L50 of 27 contigs (supplementary table S2, Supplementary Material online). The longest contig has a length of 91.5 Mbp and the GC content of the total assembly is 40.8%. Completeness analyses of three different Busco datasets, namely of the clades *Cetartiodactyla*, *Laurasiatheria*, and *Mammalia*, returned estimates of 83.4%, 90.5%, and 91.2% complete core gene sets, respectively. Repeatmasking identified a total repeat coverage of 41.8% mainly composed of retroelements (38.1%) (supplementary table S3, Supplementary Material online). The annotation of the fin whale using the transcriptome data of the minke whale (Yim et al. 2014) resulted in 17,307 complete transcripts. A functional annotation using Interprocan v5 (Jones et al. 2014) allocated potential functions for 17,152 genes, corresponding to more than 99% of all found transcripts.

Demography

We modeled the demographic history of the North Atlantic fin whale population using Stairway plot v2 (Liu and Fu 2020) which, based on the folded site frequency spectrum (fSFS, supplementary fig. S2, Supplementary Material online), estimates changes in the effective population size (N_e) over time (fig. 1). Changes in N_e over the past 800 years follow a similar trajectory for the combined number of individuals as well as for the three cohorts separately suggesting a slow and steady decline for most of the

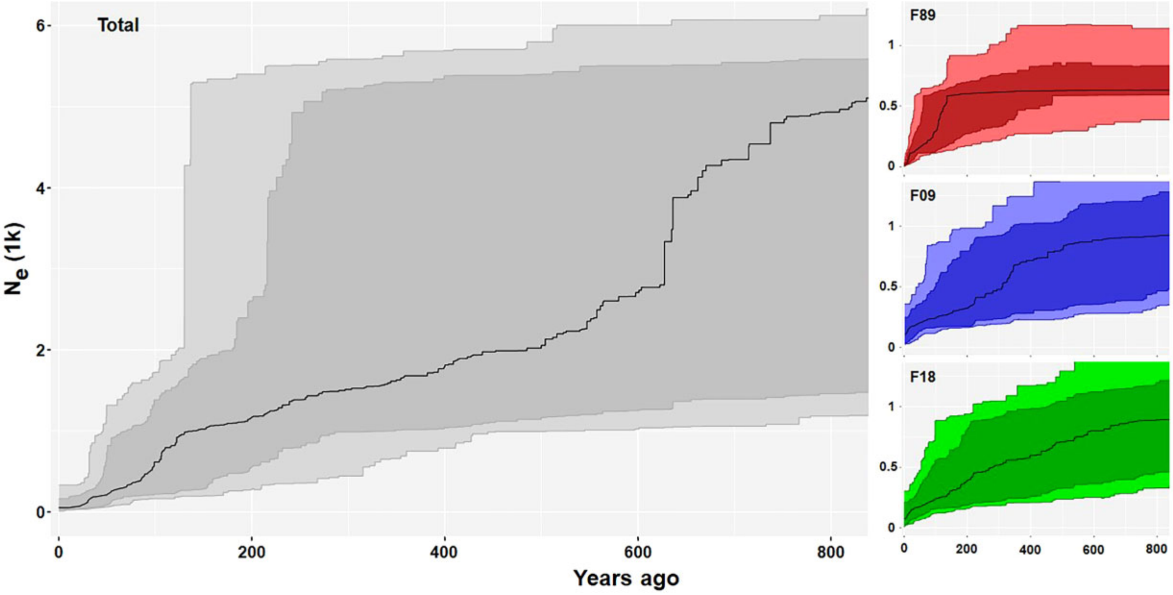

Fig. 1. Changes in N_e over the last 800 years for all analyzed fin whales combined (total, gray), or for the three fin whale cohorts 1989 (F89, red), 2009 (F09, blue) and 2018 (F18, green) separately, estimated by Stairway Plot v2 (Liu and Fu 2020). Plots were scaled using a mutation rate of 1.54×10^{-9} per site per generation and a generation time of 25.9 years. All models show a wide variety of signals, including a steep reduction in N_e around 100–150 years ago (upper 2.5% confidence interval) or only a minor gradual decline over the past 800 years (lower 2.5% confidence interval). A bottleneck pattern is more prominent in the 1989 cohort, whereas a declining pattern is more prominent in the more recent cohorts.

modeled time period. All models depict a wide variety of patterns with some showing a steep short-lived drop of 80% ~100–150 years ago while others show only minor changes. The drop in N_e is more prominent in the estimation of the combined data and the 1989 cohort compared with the estimations of the two more recent cohorts of 2009 and 2018. Nevertheless, signals of population reduction are obvious in the confidence intervals of all models. To verify these results, we simulated fSFS given a wide range of demographic scenarios using SLiM (supplementary fig. S3, Supplementary Material online, Haller and Messer 2019). Log-likelihoods of observing the empirical fSFS given one of the simulated fSFS (supplementary fig. S4, Supplementary Material online) were then calculated for each scenario and revealed that a severe population reduction leads to a more similar fSFS compared with scenarios without such an event (supplementary table S4, Supplementary Material online). Doing so also revealed that migration from a non-affected population to a population with bottleneck always weakened the performance of the respective simulation.

In addition, a pairwise sequentially Markovian coalescent (PSMC) analysis (Li and Durbin 2011) was used to model changes of N_e between 1 million (Mya) and 10,000 (kya) years ago (supplementary fig. S5, Supplementary Material online). Similar to Árnason et al. (2018), the population size first decreased over a period of 1–300 kya, then increased between 300 kya and 200 kya, before decreasing again slightly. However, the variety of patterns recorded among different individuals suggest that the demographic past could have been more complex because so6me individuals indicate more stable population trajectories compared with others.

Heterozygosity and Genetic Diversity

Genetic diversity for the study population and for other baleen whale species was estimated by genome-wide heterozygosity (He), nucleotide diversity (π), Tajima's D, and Watterson's Θ (table 1). Mean levels of heterozygosity within the fin whale population differed significantly (ANOVA f(2) = 6.1, $P = 0.005$) between the three cohorts, equaling, respectively, 0.08%, 0.09%, and 0.1% (fig. 2B). The variance within the population decreased over the three sampling periods from 2.1e−4 in 1989 to 4.6e−5 in 2018. The observed nucleotide diversity π equaled respectively 0.216, 0.23, and 0.23, whereas Tajima's D equaled 0.036, 0.036, and 0.029.

Compared with other whales, our combined fin whale data set mapped to the bowhead whale identified an average genome-wide heterozygosity of 0.07% while other baleen whales were found to have higher or lower proportions (fig. 2A). We found lower genome-wide heterozygosity in the humpback whale (0.05%), the sei whale (0.05%), and the gray whale (0.03%). In contrast, the blue whale (0.12%) and the North Atlantic right whale (0.14%) exhibited higher levels of heterozygosity.

Inbreeding and ROH

Genome-wide signs of inbreeding were studied by two different approaches. First, we measured inbreeding in the three cohorts by comparing the numbers of expected

Table 1. Statistics for Genetic Diversity, Inbreeding, and Mutational Load Inferred for Two Different Whole-Genome Data Sets, One Including 51 Fin Whale Genomes and One Including Different Available Baleen Whale Genomes.

Statistic	Cohort			Species					
	F89	F09	F18	Fin	Sei	Blue	Gray	Humpback	NA right
Mean He	0.079	0.088	0.102	0.068	0.045	0.118	0.026	0.05	0.141
Mean (π)	0.252	0.255	0.261	0.099	—	—	—	—	—
Tajima's *D*	0.036	0.023	0.029	−0.008	—	—	—	—	—
Watterson Θ	0.216	0.232	0.232	0.107	—	—	—	—	—
F_H	−0.007	−0.038	−0.07	—	—	—	—	—	—
F_{ROH} (>1 Mbp)	0.019	0.004	0.006	0.081	0.046	0.195	0.027	0.050	0.427
# LoF Mutations	258	256	289	1108	1443	1685	1792	228	1094
Mutational Load	$1.6e^{-04}$	$1.1e^{-04}$	$1.4e^{-04}$	$2e^{-04}$	$1.8e^{-04}$	$1.9e^{-04}$	$2.3e^{-04}$	$2.1e^{-04}$	$2e^{-04}$
# He LoF	157	157	187	119	141	235	77	32	344
# Ho LoF	101	100	102	989	1302	1450	1715	196	750

He, heterozygosity.
Ho, homozygosity.
π, nucleotide diversity.
F_H, inbreeding coefficient following Kardos et al. (2015).
F_{ROH}(>1 Mbp), inbreeding factor based on runs of homozygosity over 1 Mbp in percent.
# LoF, mean number of loss of function mutations.
NA right, North Atlantic right whale.
The 51 fin whale individuals are further differentiated into three cohorts based on their sampling year (F89 = 1989, F09 = 2009, and F18 = 2018). Some statistics are not applicable (—) for all species because most other baleen whales were represented by a single individual per species.

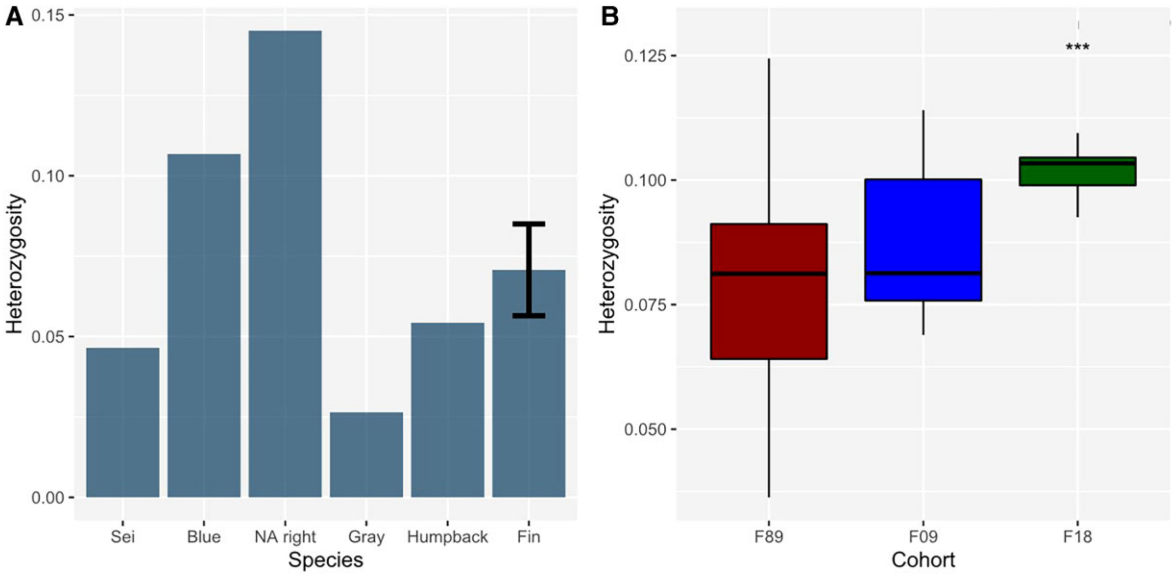

Fig. 2. Genome-wide heterozygosity, (*y*-axis) in percent of heterozygous sides in the SNP and SNV data sets, respectively. (*A*) Heterozygosity within the baleen whale SNV data set using the bowhead whales as a reference. Fin whales show a moderate heterozygosity of around 0.07%. The blue whale and North Atlantic right whale individuals had a higher He (0.11% and 0.15%), whereas sei, gray, and humpback whales were less heterozygous (0.05%, 0.03%, and 0.05%). (*B*) Box plot of the He distribution between three fin whale cohorts sampled in 1989 (red), 2009 (blue), and 2018 (green), respectively. A slight increase in He over the three cohorts was observed, with individuals sampled in 2018 differing significantly ($P = 0.0051$) from the 1989 and 2009 cohorts.

heterozygous sites against the observed number (inbreeding factor F_H) using SambaR's "calckinship" function, following the definition of Kardos et al. (2015) (table 1, supplementary fig. S6, Supplementary Material online). Inbreeding factors F_H were slightly negative in all three fin whale cohorts. Although a F_H of −0.007 in 1989 identifies a nearly expected number of homozygous genotypes, a decrease to −0.038 in 2009 and to −0.070 in 2018 suggests a slight excess in heterozygous genotypes compared with expected values.

We calculated inbreeding factors (F_{ROH}) based on ROH as the coverage of runs exceeding a defined size cutoff, beginning with 100 kbp and increasing stepwise to 1 Mbp (fig. 3*B*). A similar gradually decreasing pattern of inbreeding factors in the defined length bins was identified in all three cohorts, beginning with an average of ~3% in the 100–200 kbp bin and declining to an average of ~1% in the >1 Mbp bin. Apart from this general pattern, four outlier individuals were noticed in the 1989 cohort, featuring more or less ROH in most of the bin. Furthermore,

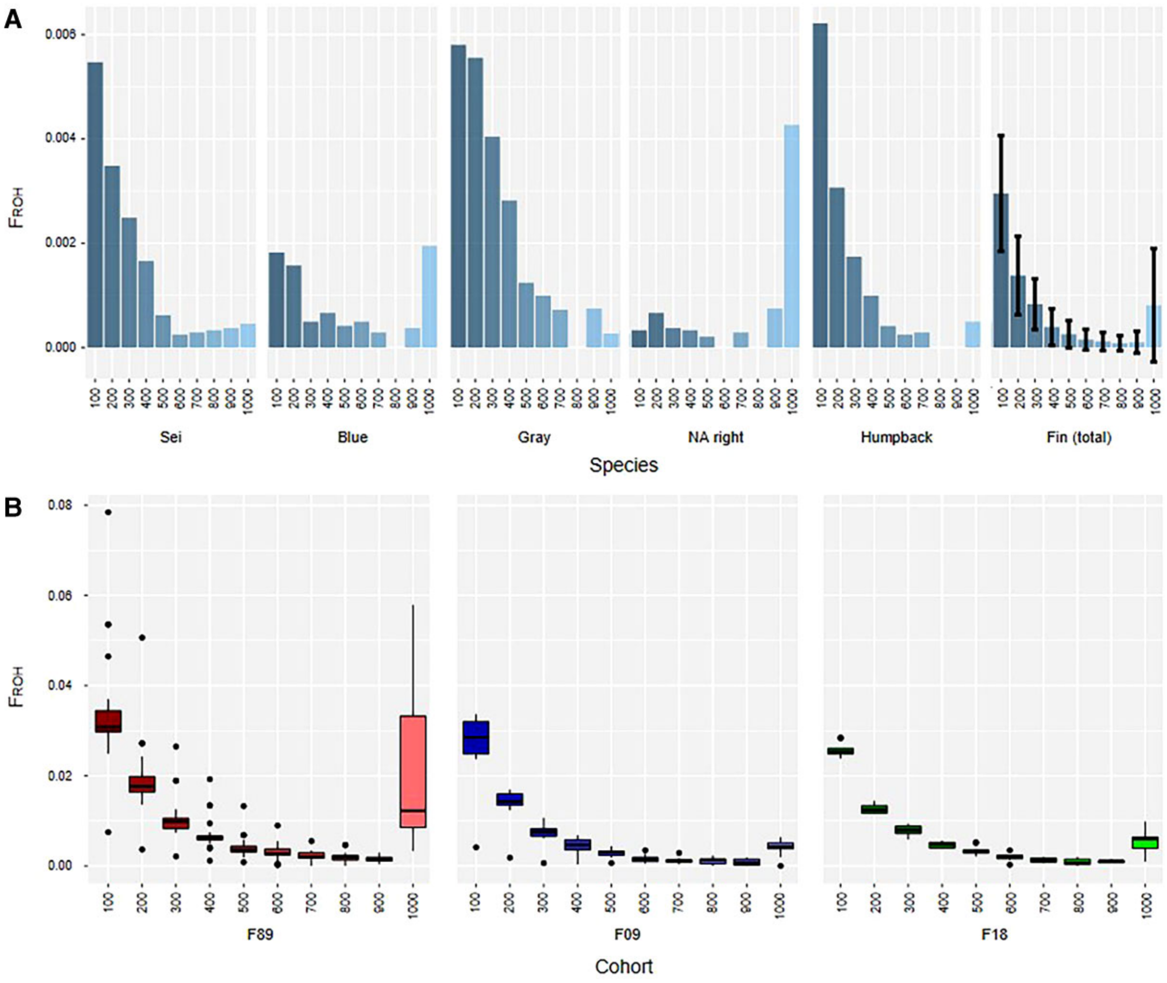

Fig. 3. Inbreeding factors (F_{ROH}) based on the genome coverage of run of homozygosity (ROH) between different minimal lengths cutoffs of ROH: 100 kbp to 1 Mbp (*x*-axis in 100 kbp steps). (*A*) Comparison of F_{ROH} among genomes of different baleen whale species and the 51 fin whales. While sei, gray and humpback whale have similar patterns of gradually decreasing inbreeding coefficients in their ROH length bins from 0.6% to <0.1%, varying patterns of high F_{ROH} (~0.2–0.4%) in the >1 Mbp bin and low F_{ROH} (<0.1–0.2%) in the 100 kbp bin were found in the genomes of blue and North Atlantic right whale indicating more recent inbreeding events. The fin whale population shows an overall similar pattern to the sei, gray and humpback whale but with potentially higher inbreeding coefficients on the >1 Mbp length bin in some of the individuals (~0.2%). (*B*) Comparison of F_{ROH} between the three different fin whale cohorts 1989 (red), 2009 (blue), and 2018 (green). We found four outlier individuals with higher or lower inbreeding coefficients and a significantly higher amount of long >1 Mbp ROH (F_{ROH} up to 0.6%) in the 1989 cohort. In the other two cohorts, only the same gradually decreasing pattern of inbreeding coefficients was identified.

significantly more ROH of the longest category (>1 Mbp) were found within the 1989 cohort compared to both other cohorts (table 1, supplementary fig. S8, Supplementary Material online), indicating that some individuals within this cohort experienced more recent inbreeding.

When compared with other baleen whales (fig. 3A), fin whales featured the same gradual decrease as identified for the individual cohorts (total F_{ROH}: 0.7%). Similar distributions but with generally higher inbreeding coefficients were noticed for the sei (total F_{ROH}: 1.5%), gray (total F_{ROH}: 2.2%), and humpback whale (total F_{ROH}: 1.3%). In the blue and North Atlantic right whales, however, divergent distributions without this gradual decrease were found. In those individuals, low or lowered numbers of short ROH were recorded, whereas long ROH (>1 Mbp) were much more frequent (table 1). Especially in the North Atlantic right whale, long ROH accounted for more than half (0.4%) of the total inbreeding coefficient of 0.7%. Long ROH in the blue whale genome made up 0.2% of the total 0.8% F_{ROH}. Furthermore, we found a negative correlation trend between the total F_{ROH} coefficients and genome-wide heterozygosity (fig. 5A) within the combined data set. Inbreeding factors F_H and genome-wide heterozygosity showed a slight positive correlation trend.

Mutational Load

Mutations with a potentially negative fitness impact were identified by annotating our single nucleotide

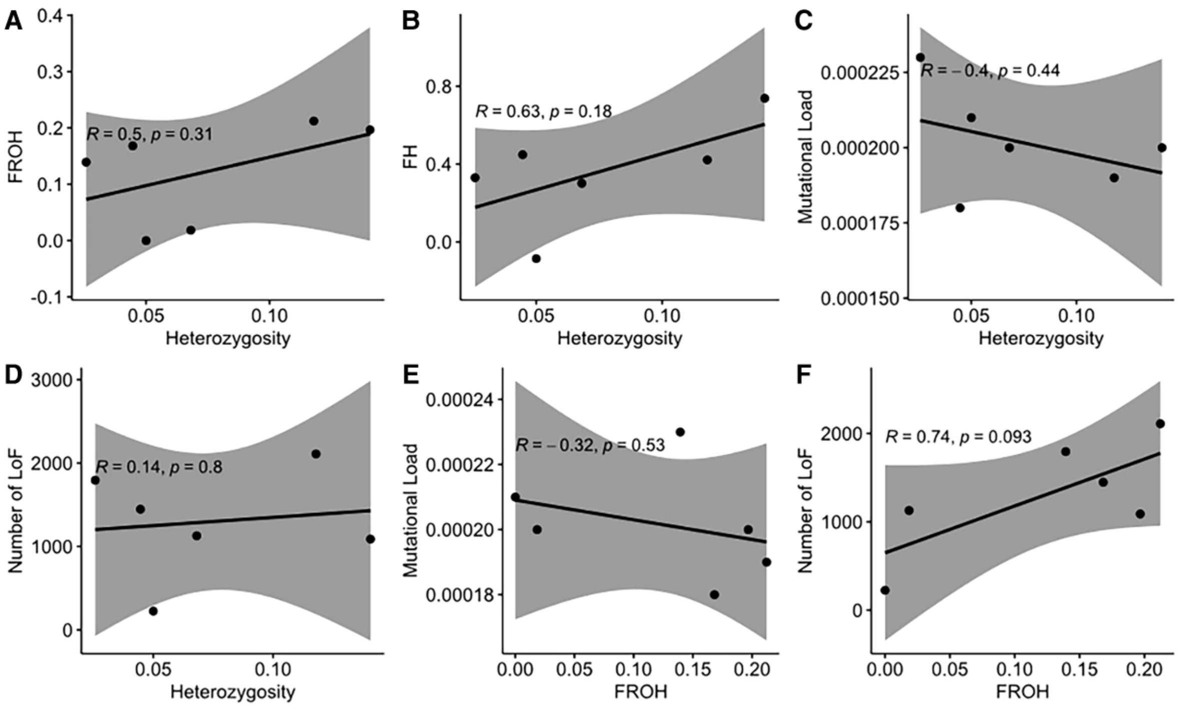

Fig. 5. Cross correlation analysis between genome-wide heterozygosity, inbreeding and mutational load showing the potential correlations between heterozygosity and F_{ROH} as well as F_H (*A* and *B*) and between the relative frequency of LoF mutations (*C* and *E*) or total number of LoF mutations (*D* and *F*) against heterozygosity or F_{ROH}. We identified a lack of correlations between nearly all parameters. Only heterozygosity and F_{ROH} have a non-significant negative trend towards each other as indicated by a $R = -0.81$ and $P = 0.053$. Neither the relative frequency nor the total number of LoF mutations showed any such trends.

polymorphism (SNP) and single nucleotide variants (SNV) datasets based on the genome annotation of this study or from the bowhead whale genome resource (Keane et al. 2015). Using SNPeff v4.3 (Cingolani et al. 2012), an average of 1.26% of our fin whale datasets and 2.4% of our baleen whale dataset were labeled as functional mutations and sorted into the three categories: *synonymous*, *missense and loss of function* (LoF). Based on these annotations, we found on average 0.69% synonymous mutations, 0.58% missense mutations, and 0.01% LoF mutations in the fin whales SNP data while we identified 1.37% synonymous, 0.99% missense, and 0.02% LoF mutations in the SNV data of all baleen whales (fig. 4*A*). Between the different cohorts, no significant differences were observed in the three functional categories (fig. 4*B*). Fin whales from the 1989 cohort featured the most annotated mutations in every category, but individuals from the 2018 cohort possessed a slightly higher mean number of LoF mutations. All three cohorts had a similar number of variants in a homozygous state (∼100), yet the 2018 cohort showed on average 30 heterozygous LoF mutations more compared with the other two.

Among other baleen whale species, only minor differences in the mutational load were identified. However, the North Atlantic right whale seems to have a slightly increased proportion of LoF mutations respective to the proportions of other categories, but none of these differences were significant (fig. 4*A*, *C*, and *E*).

Finally, no significant correlations were observed between the mutational load or the total number of LoF mutations to either genome-wide heterozygosity or the inbreeding coefficient based on ROH (fig. 5*C*–*F*). The relationship between mutational load and heterozygosity shows a more negative trend, and the relationship between mutational load and inbreeding shows a more positive trend.

Population Structures

The two conducted population structure analyses identified no sub-structures within the sampled fin whale individuals (fig. 6). The admixture analysis conducted with the Lea package (Frichot et al. 2015) produced random signals of admixture that affected all individuals regardless of the assumed K. The principal coordinate analysis (PCoA) found only one cluster (PC1 depicts 3.1% variance, PC2 3.1%).

Discussion

In this study, we analyzed the genome-wide diversity, inbreeding, mutational load and the demographic history of a North Atlantic fin whale population and other baleen whales that might have been affected by large-scale whaling in the past. Modeling the recent demographic history of the fin whale population using the site frequency spectrum, we were able to quantify the

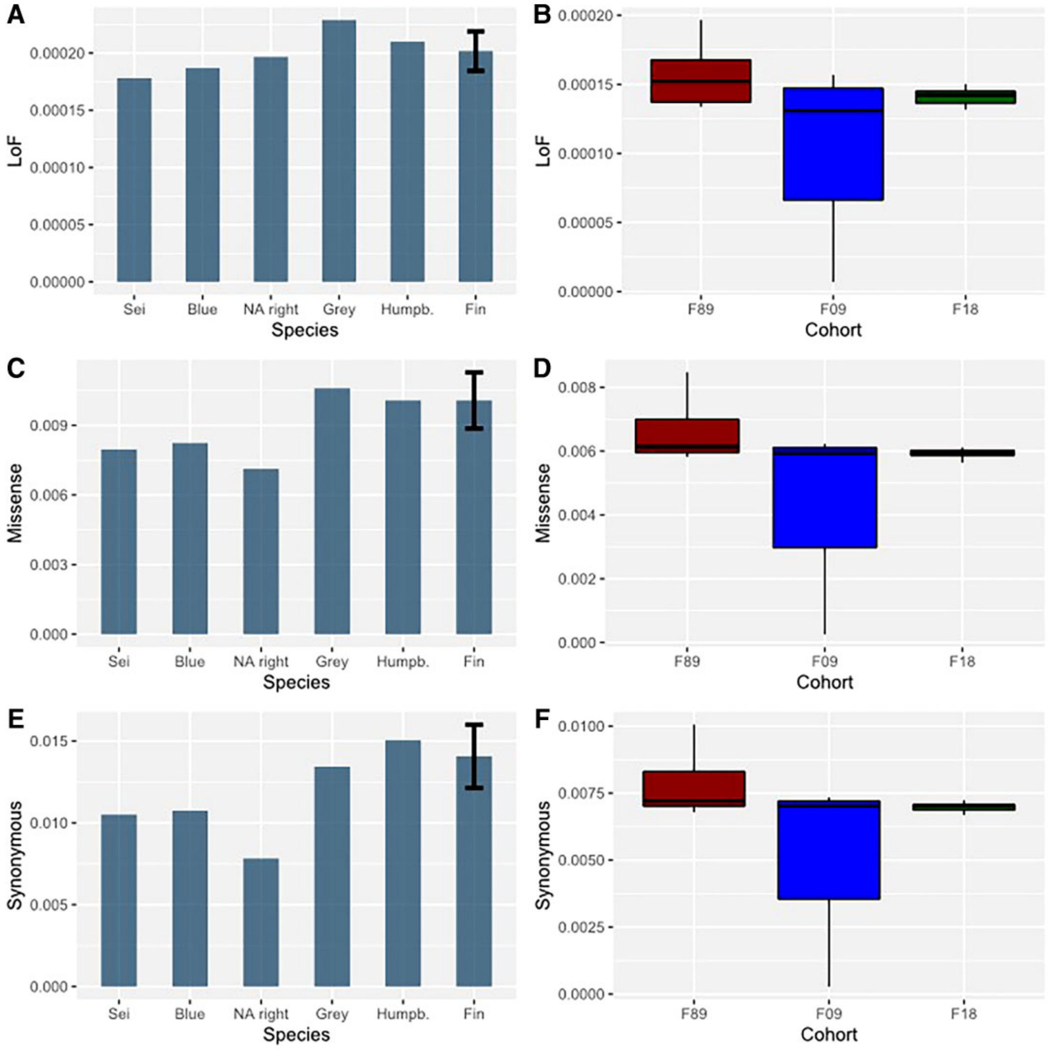

Fig. 4. Abundances of three categories of functional mutations (LoF, missense and synonymous), measured as their relative proportion compared with the total number of SNPs or SNVs, respectively. Within the baleen whale data set (*A*, *C* and *E*), fin whales show a relatively high abundance of mutations in every category. By contrast, the North Atlantic right whale has a comparable high number of LoF mutations relative to the respective abundances of missense and synonymous mutations. No significant differences were found in the 51 fin whale genomes between the three sampling years 1989 (red), 2009 (blue), and 2018 (green) (*B*, *D*, and *F*). Fin whales from 1989 always have the most mutations in every category while whales from 2009 always have the least number of mutations. The overall variation in the 2018 cohort was, in general, lower compared with the other two cohorts. Yet, all cohorts had a similar medium number of mutations in every category.

reduction of the population a century ago. Thus, the exploitation of fin whales left a noticeable signature in their genomes that coincides with maximum hunting pressure on the species (Tønnessen and Johnsen 1982). It is uncertain if this signature is characteristic for the North Atlantic population, or, more likely, is found in all fin whales due to the worldwide exploitation of this species in the past (Tønnessen and Johnsen 1982; Aguilar and García-Vernet 2017).

Despite the genomic and documented impact of whaling on the effective population size (Tønnessen and Johnsen 1982), North Atlantic fin whales do not show signs of genomic consequences that would affect their overall genetic fitness. The population features a moderate level of heterozygosity and neither excessively long ROH nor a pronounced excess of loss of function mutations when compared with other baleen whales. Instead, found genetic consequences are relatively weaker compared with other baleen whales that were proportionally more affected by whaling like the blue whale (Tønnessen and Johnsen 1982) and are comparable with baleen whales that were hunted on a similar or lower scope like the sei whale and the humpback whale (Tønnessen and Johnsen 1982), respectively. Other genomic studies addressing more threatened or potentially extinct species and populations such as the Grauer's gorilla, the Scottish killer whale, or the Malay Peninsula rhinoceros (van der Valk et al. 2019a; Foote et al. 2021; von Seth et al. 2021) show more pronounced genomic consequences, such as substantial genome coverage of ROH longer than 1 Mbp and a significant increase of loss of function mutations. Therefore, our study does not support a molecular threat

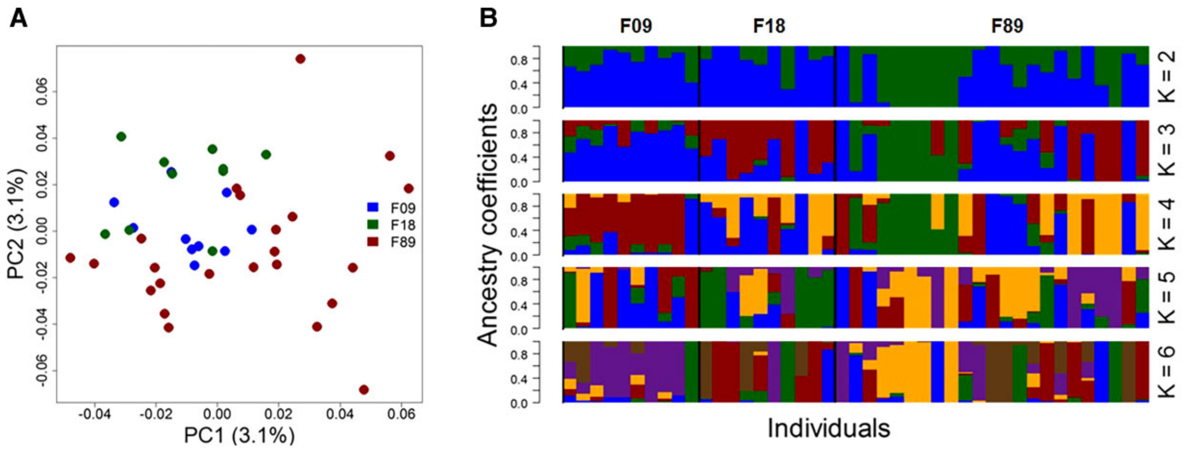

Fig. 6. Population structure analyses of fin whales sampled in Icelandic waters in 1989 (red), 2009 (blue), and 2018 (green), respectively. (*A*) PCoA identified only one major fin whale population. (*B*) The admixture-like analysis (colors indicate clusters inferred by the algorithm) resulted in no clear structure, indicating free exchange of genetic material in this population over all three cohorts.

of the North Atlantic fin whale population. Instead, we suspect the levels of genetic diversity, inbreeding, and mutational load to be the result of a long-lasting demographic pattern with minor, natural fluctuations with limited impact on the genomes of the population.

The fact that the population bottleneck inflicted by whaling was short-lived and relatively recently might explain the lack of negative genetic consequences in fin whales. In theory, negative effects could still appear in later generations, because, for example, the mutational load is expected to not increase drastically unless the bottleneck is persistent (Teixeira and Huber 2021). Given the expected generation time of about 26 years for fin whales (Taylor et al. 2007), ~4–6 generations have passed since whaling peaked in 1915. However, population-genetic theory and empirical findings suggest that a bottleneck of a few generations can cause long ROH and decreased levels of heterozygosity (Nei et al. 1975; Gurgul et al. 2016). Both effects were not observed in the presented sampling. Instead, we observed a significant increase in heterozygosity over the studied 30-year period and only few regions of long ROH.

An alternative explanation might be a genetic exchange with other populations or even species, which may be supported by signals of no reduction in the demographic analysis (fig. 1) and by the occurrence of outlier individuals in the ROH analysis of the 1989 cohort (fig. 3). Such a genetic exchange is often desired in conservation management plans by adding individuals from a stable population to a threatened one and is often referred to as a “genetic rescue” (Frankham 2015). Seven stocks of fin whales are currently assumed by the North Atlantic Marine Mammal Commission for the North Atlantic (NAMMCO 2005) and at least two genetically and morphologically distinct populations have been verified (Bérubé et al. 1998). It is possible that genetic exchange between one of those populations and the Icelandic fin whales weakened the genetic consequences of extensive hunting. In addition, genetic exchange between different ocean basins is evident from the analyses of mitogenome and SNP data (Archer et al. 2019). Although genetic exchange cannot be excluded, the relative genetic uniformity between most fin whale individuals suggests that genetic exchange between different populations is rare and may not explain the overall lack of negative genetic consequences in the fin whale population. Furthermore, our demographic simulations resulted in lower model performance when including migration from a non-affected population to a bottleneck population, which either shows that all fin whale populations were impacted by whaling or that migration happens on a small scope.

Introgression from blue whale genomes might be another possibility and may also have contributed to the genetic diversity of North Atlantic fin whales. However, blue whales are potentially more affected by census size depletion and show more signs of genetic consequences compared to the here analyzed fin whales (figs. 2–4). Furthermore, it seems that introgression between both species is unidirectional from fin to blue whale (Jossey et al. 2021; Pampoulie et al. 2021).

Despite the substantial impact of whaling on the effective population size of North Atlantic fin whales, the apparent lack of other genomic consequences challenges the common concern of a fatal over-exploration of fin whales by 19th and 20th century whaling. Instead, it could be possible that the bottleneck inflicted by Icelandic whaling never reached a duration or scale that would have triggered widespread genomic changes. Iceland was involved in industrial whaling for three decades between 1883 and 1915 (Tønnessen and Johnsen 1982). During this time, about 60,000 captured baleen whales have been reported in the complete Northern Atlantic. It can be expected that a major proportion of these catches were fin whales, however, estimating total numbers is problematic due to limited documentation during this time. In any case, by 1915, whaling became unprofitable in Icelandic waters as catch rates decreased and prices of whale oil inflated, majorly influenced by the increased availability and use of mineral oil. Iceland enforced a complete, two decades long ban on whaling at that time, before taking up new activities on a smaller scope. Owing to the uncertainty of fin whale catch numbers, the period without whaling, and because catch rates stayed on a relatively low and constant level since 1948 (Árnason 1981), it is possible that the bottleneck was not as severe compared to other areas of the world like in the waters of Finnmark, Norway, or compared with other whale species such as the blue whale (Tønnessen and Johnsen 1982).

By contrast, another whale species, the North Atlantic right whale, seems to show negative genetic effects from population depletion. Our genomic analyses revealed extensive ROH coverage and disproportionate, high levels of LoF mutations, notwithstanding high levels of heterozygosity. It is feared that there are <500 individuals remaining worldwide, which led to the classification as "critically endangered" on the IUCN red list (Cooke 2020). In addition, records of pre-industrial whaling (Tønnessen and Johnsen 1982), an increased anthropogenic mortality (Kraus et al. 2005) and slow reproduction rates (Browning et al. 2010) exist, indicating a long and persistent bottleneck for the species. High levels of genetic diversity contradicts previous findings based on microsatellites and mitochondrial marker but could potentially explained with what was previously reported by Frasier et al. (2013). The authors suggest that mating of genetically dissimilar individuals due to postcopulatory selection of gametes can lead to more heterozygous individuals compared to what would be expected by random mating. Although we cannot make definitive statements about this based on a single specimen, our finding of putatively substantial inbreeding opposes this assumption of non-random mating and furthermore contradicts the common assumption of the existence of a reciprocal relationship between population sizes, heterozygosity, inbreeding and mutational load which would lead to an "extinction vortex" as a consequence of a bottleneck that persisted for long periods of time (Fagan and Holmes 2006; Blomqvist et al. 2010; Teixeira and Huber 2021).

A correlation test of these parameters in all here analyzed baleen whales indicated no significant relationships between either of those factors. A nearly significant negative correlation was only identified between genome-wide heterozygosity and ROH, which is expected given their linked relationship. Nonetheless, the apparent absence of clear-cut relationships between effective population size, heterozygosity, ROH, and mutation load and IUCN status is consistent with previously reported findings. There is, for example, no significant correlation between the IUCN red list status of a population and their levels of inbreeding (measured by ROH) or genome-wide heterozygosity (Brüniche-Olsen et al. 2018) potentially induced by, for example, non-random mating of genetically dissimilar individuals as described in Frasier et al. (2013). There are also numerous examples of small populations with low genetic diversity and higher inbreeding that do not suffer from deleterious mutations due to purging, which further complicates the picture (Robinson et al. 2018; van der Valk et al. 2019b; Ochoa and Gibbs 2021). This suggests that if a population is heavily reduced, it might resist the reduction of genetic diversity or might resist the consequences of low genetic diversity. Therefore, we propose, similar to previous discussions (Teixeira and Huber 2021), that the reciprocal relationship between heterozygosity, inbreeding, and mutational load is not as direct as previously assumed and that measuring only one of those parameters could misjudge the actual level of endangerment of a population.

In the case of the analysis of the single genome of the North Atlantic right whale, this implies that their potentially high genetic diversity may not indicate a lowered risk of extinction. Instead and discussed earlier (van der Valk et al. 2019a), their increased levels of inbreeding and mutational load combined with a higher heterozygosity might even increase the risk of extinction disproportionally, because the potential higher number of deleterious mutations could become fixed more rapidly due to the emergence of ROH (van der Valk et al. 2019b; Teixeira and Huber 2021). To further evaluate the genetic risk in this species, population-genomic studies, like those presented here for the fin whale, are necessary to evaluate their population on a genomic level for targeted conservation efforts.

Conclusion

Genome data of North Atlantic fin whales made it possible to assess the impact of whaling on the genetic diversity of a baleen whale population. Demographic analyses confirmed, consistent with historical records, that the population experienced a substantial reduction in its census size a century ago. Despite the decimation of their population and relatively low levels of heterozygosity compared to other whales, fin whales have a stable or even slightly increasing genome-wide diversity over time. In addition, there is no evidence for increased inbreeding or mutational load suggesting that the bottleneck, caused by whaling, had less impact on the genotype of the species as previously feared.

By contrast, analyses of other baleen whales revealed that the most threatened baleen whale species, the North Atlantic right whale, has relatively high levels of inbreeding and mutational load despite their potentially high genetic diversity. This calls for population-level genome-sequencing efforts for other baleen whales to enable a comprehensive conservation-genomic assessment and targeted conservation strategies.

Materials and Methods

Sampling, DNA Isolation, and Sequencing

A total of 51 tissue samples from individual fin whales were collected during fisheries operations in Icelandic waters in 1989, 2009, and 2018 (supplementary fig. S1, Supplementary Material online). The operation in 1989 was conducted under scientific research permit of the Icelandic Ministry of Food, Agriculture and Fisheries, operating in the years 1986–1989. The sampling in the years 2009–2018 was done during commercial fisheries operations licensed by the Icelandic Ministry of Food, Agriculture and Fisheries. Individual licenses are available on request.

Tissue samples were stored in 96% ethanol at −20°C and DNA was extracted from ∼20–50 mg tissue using a standard phenol-chloroform-isoamylalcohol protocol (Sambrook and Russell 2006). DNA libraries were prepared and sequenced by SciLifeLab, Stockholm, Sweden, or by Novogene, Cambridge, and United Kingdom. SciLifeLab and Novogene libraries were generated with the Rubicon ThruPLEX DNA-seq kit and the NEBNext DNA Library Prep kit, respectively, both according to manufacturer's recommendations and using 350 bp insert size. Illumina short read sequencing was performed using the Illumina NovaSeq 6000 platform to produce 10-fold sequence coverage or ∼24 Gbp of 150 bp paired-end reads per individual. In addition, a single 10× Genomics Chromium library was compiled and sequenced by SciLifeLab, yielding 487,342,309 paired/linked 150 bp Illumina short reads (∼30-fold coverage).

De novo assembly

A de novo genome of the fin whale was assembled using the linked short reads sequenced with the 10× Chromium technology. We used Supernova v2.1.1 (Weisenfeld et al. 2017) to construct pseudo-haplotype assemblies and evaluated their properties using Quast v5.0.2 (Mikheenko et al. 2018). Additional scaffolding and correction steps were performed using Arcs v1.1.1 (Yeo et al. 2018) and Tigmint v1.1.2 (Jackman et al. 2018) which led to no substantial improvements, and we proceed using the best raw pseudo-haplotype assembly. The assembly was then assessed for coverage distribution (Qualimap v2.2.2, Okonechnikov et al. 2016) and gene set completeness (Busco v4.1.1, Seppey et al. 2019).

Repeat and Genome Annotation

We screened the assembly for repetitive sequences using RepeatModeler v2 (www.repeatmasker.org) and merged found repeats with the *Cetartiodactyla* database from RepBase (Jurka et al. 2005). The merged data set was used to mask repeats in our assembly using RepeatMasker v4.1 (www.repeatmasker.org). Evidence and homology-based gene annotation was performed with the Maker v2.31 pipeline (Holt and Yandell 2011) using data sets from the northern minke whale, *Balaenoptera acutorostrata* (Yim et al. 2014). Furthermore, genes were predicted using Augustus v3.2.2 (Stanke et al. 2006) and GeneMark-ES v4 (Lomsadze et al. 2005) as implemented in Maker. Finally, we annotated gene functions to the predicted protein sequences using Interproscan v5 (Jones et al. 2014) with default parameters.

Read Mapping and SNP Calling/Filtering

Short read sequences were examined using FastQC v0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed for read quality with Fastp (Chen et al. 2018) and for adapter sequences with AdapterRemoval v2 (Schubert et al. 2016). Trimmed reads were mapped to the de novo assembled fin whale genome using BWA-mem v0.7.17-r1188 (http://bio-bwa.sourceforge.net). Potential duplicates were removed, and read-groups were added using the Picard v2.21.2-0 toolkit (https://broadinstitute.github.io/picard/). Genotypes including multi-variant and monomorphic sites were called in a combined approach including all mapping files as well as individually per single mapping file to account for different needs of downstream analyses. This was done with BCFtools v1.12 mpileup and BCFtools v1.12 call (Danecek et al. 2021) using the "-m" or "-c" flags respectively applying a minimal mapping- and base-quality cutoffs of 30 using the flags "-q" and "-Q". BCFtools v1.12 filter (Danecek et al. 2021) was used to exclude indels, sites with divergent read coverage (>3-fold and <0.3-fold of the expected mean coverage) and sites with more than 5% missing data. In the case of the combined data set, VCFtools v0.1.16 (Danecek et al. 2011) was used to remove multi-variants sites as well to retrieve SNPs. In addition, for the combined data set, Plink v1.90p (Chang et al. 2015) was used with "–indep-pairwise 1,000 kb 1 0.9" parameters to remove sites in potential linkage disequilibrium. Furthermore, putatively related individuals were removed after an identity-by-descent test using the "–genome" function of Plink v1.90p (Chang et al. 2015) applying an pi_-hat cutoff of 0.2. The final combined dataset consists of 966,242,959 genotypes and 7,022,898 SNPs whilst the individually called data sets contain between 976 Mio and 1 Bio genotypes. To compare our findings against other whale species, these steps were repeated by mapping raw reads of five different baleen whale species (Árnason et al. 2018) sequenced with the same sequencing platform like used in this study. All five genome data sets and the here presented fin whale genomes were mapped against the bowhead whale, *Balaena mysticetus*, reference genome (Keane et al. 2015). To ensure comparability, we repeated each step, starting from quality filtering of raw reads up to the filtering of single nucleotide variances (SNVs), all with equal filter parameters. This second combined data set consists of 56 individuals, six baleen whale species, 469,467,070 genotypes, and 14,857,736 SNVs. Individually called data sets contain between 115 Mio and 1 Bio genotypes.

Genetic Structure and Population Differentiation Analysis

We divided the fin whale samples into multiple cohorts based on their capture years: 1989, 2009, and 2018.

Population structure analyses were then performed on the combined fin whale SNP data set that was further thinned randomly (one SNP per 1 kbp) with VCFtools v0.1.16 thin (Danecek et al. 2011) to reduce the computational load of the following steps. The R-package SambaR (de Jong et al. 2021) was used to filter out individuals with more than 5% missing data as well as SNPs with more than 10% missing data, heterozygosity excess, and a minor allele count of 1. Population-genetic analyses were performed by using SambaR's main functions "findstructure()" and "calcdistance()". Among the analyses invoked by these wrapper functions are PCoA performed with the ape-5.3 package (Paradis and Schliep 2019) and admixture analysis performed with the LEA-2.4.0 package (Frichot et al. 2015).

Genetic Diversity

Nucleotide diversity, Watterson's θ and Tajima's D estimates were generated using SambaR's main function "calcdiversity()". Genome-wide heterozygosity was inferred by counting heterozygous sites in the individual VCF files that still included multi-variant and monomorphic sites. To test for potential significant differences in genome-wide heterozygosity between the three cohorts, an ANOVA test was conducted. A fSFS was generated using the "vcf_to_sfs" tool distributed within the Popgen Pipeline Platform (Webb et al. 2021).

Demographic Analysis

Demographic history of the fin whale population was inferred based on the fSFS using the Java package Stairway plot v2 (Liu and Fu 2020) which estimates series of mutations rates over time. For all demographic estimations, we defined a mutation rate of 1.54×10^{-9} per site per generation following Tollis et al. (2019) and a generation time of 25.9 years following Taylor et al. (2007). We determined the modeled time window to the last 800 years (~30 generations) and conducted analyses based on the fSFS of the total fin whale sampling as well as on the fSFS of the individual cohorts (fig 1).

To further elucidate the role of different demographic events on the fSFS, we implemented forward-in-time Wright-Fischer simulations using SLiM v3.7 (Haller and Messer 2019) to generate expected fSFS given a certain scenario (supplementary table S4, Supplementary Material online). For each simulation, we started with a population of 50,000 individuals and three 10 Mbp long genomic elements. The mutation rate was set to 1.54×10^{-9} per site per generation and the recombination rate was defined as 1×10^{-8} per generation. Each simulation ran without any events for 100,000 generations to obtain neutral fSFS. Afterwards, demographic changes were applied as depicted in supplementary figure S3A–H, Supplementary Material online and a number of individuals were sampled equal to the number of fin whale individuals. fSFS were extracted from the resulting vcf files using the "vcf_to_sfs" tool as described above (Webb et al. 2021). Eventually, we compared different demographic simulations by calculating log-likelihoods of observing the empirical fSFS given one of the simulated fSFS using the R base function "dmultinom" (supplementary table S4, Supplementary Material online).

Additionally, we used the PSMC framework (Li and Durbin 2011) to model historical N_e further back in time (10 kya to 1 Mya) using the individual vcf files constructed with the fin whale reference genome as described before. These files were filtered as described above before inferring consensus sequences with BCFtool's vcfutils.pl. Consensus sequences were then used for the PSMC modeling using the same mutation rate and generation time as for the Stairway plot analysis (supplementary fig. S5, Supplementary Material online).

Inbreeding Estimation and ROH

Inbreeding factors based on the excess of homozygous sites were estimated for the complete fin whale dataset only as it requires assumptions about the expected number of heterozygous sites per population (supplementary fig. S6, Supplementary Material online), whereas ROHs were collected for all individual vcf files including the data of other baleen whale species. By comparing proportions of observed and expected homozygous sites using SambaR's "clackinship" function, inbreeding coefficients for all three cohorts were gathered (F_H, supplementary fig. S6, Supplementary Material online, Kardos et al. 2015). ROHs were identified with Darwindow (https://github.com/mennodejong1986/Darwindow), which finds ROH per individual with a sliding window approach. ROHs were detected using a sliding window size of 10 kbp, a heterozygosity threshold of 0.2%, and a minimal window number of 10. Excluded from the analysis were scaffolds with a size below 3 Mbp. Found ROH were subsequently sorted into different length bins ranging from 100 kbp to over 1 Mbp with a step size of 100 kbp. Detailed graphs with bins from 100 kbp to over 4 Mbp are depicted in supplementary figure S7, Supplementary Material online. Individual inbreeding coefficients per bin were calculated from the extent of ROH spanning the respective reference genome (F_{ROH}, after McQuillan et al. 2012) and an ANOVA test was conducted to find significant differences between F_{ROH} of the different fin whale cohorts (supplementary fig. S8, Supplementary Material online). Potential correlations were estimated between genome-wide HE and either F_{ROH} or F_H in R using a Pearson correlation test.

Mutational Load

Mutational load was inferred based on the functional annotation of variants. SNPs and SNVs were assigned with potential functional categories with SNPeff v4.3 (Cingolani et al. 2012) using the annotation generated in this study and the annotation generated by Keane et al. (2015). To get the total number of variants, individual vcf files were filtered as described above before annotating them with SNPeff using default parameters. Resulting functionally assigned variants were sorted into the categories *synonymous*, *missense*, and *loss of function* (LoF), and normalized using the total number of variants per individual. Mutational load was defined as the proportion of LoF mutation compared with the respective total counts of variances. The numbers of LoF mutations were furthermore differentiated between heterozygous and homozygous variants to estimate which proportion might actually affect the fitness of the individual. Finally, potential correlations between either the total number of LoF mutations or the relative abundance (mutational load) were inferred from genome-wide HE and F_{ROH} applying a standard Pearson correlation test in R.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We are grateful to the Marine and Freshwater Research Institute, Reykjavik, Iceland, for their sampling efforts, for providing tissue samples, and for the assistance with the sequencing process. Furthermore, we thank Sven Winter for his support in general lab work and bioinformatic analyses.

Funding

The present study was funded by the Centre for Translational Biodiversity Genomics (LOEWE-TBG) through the program "LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts, as well as the Carl Tryyger foundation, the Royal Physiographic Society, Lund, and the Erik Philip-Sörensen's Foundation

Author Contributions

M.W., U.A., and A.J. conceived and designed the study, M.W., M.J., U.A., and A.J. wrote the manuscript, M.W. conducted the analyses, M.J. aided with the computational analyses of the re-sequencing data, and S.D.H. conducted the sampling.

Data Availability

Raw sequencing reads have been deposited at the National Center for Biotechnology Information under the BioProject PRJNA740292. The assembled genome sequence of the fin whale is deposited as Genome: JAHXJN000000000, BioSample: SAMN19897699. All other data needed to evaluate the conclusions of the article are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors.

References

Aguilar A, García-Vernet R. 2017. Fin Whale: *Balaenoptera physalus*. In: Würsig BG, Thewissen JGM, Kovacs KM, editors. *Encyclopedia of marine mammals*. Amsterdam: Academic Press.

Archer FI, Brownell RL Jr, Hancock-Hanser BL, Morin PA, Robertson KM, Sherman KK, Calambokidis J, Urbán RJ, Rosel PE, Mizroch SA, *et al.* 2019. Revision of fin whale *Balaenoptera physalus* (Linnaeus, 1758) subspecies using genetics. *Journal of Mammalogy*. **100**(5): 1653–1670.

Árnason Ú. 1981. Fin whales in the NE Atlantic; relationships between abundance and distribution. *Ecography*. **4**(4):245–251.

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**(4): eaap9873.

Bérubé M, Aguilar A, Dendanto D, Larsen F, Di Notarbartolo Sciara G, Sears R, Sigurjónsson J, Urban-R J, Palsbøll PJ. 1998. Population genetic structure of North Atlantic, Mediterranean Sea and Sea of Cortez fin whales, *Balaenoptera physalus* (Linnaeus 1758). Analysis of mitochondrial and nuclear loci. *Mol Ecol*. **7**(5):585–599.

Blomqvist D, Pauliny A, Larsson M, Flodin L-A. 2010. Trapped in the extinction vortex? Strong genetic effects in a declining vertebrate population. *BMC Evol Biol*. **10**:33.

Booy G, Hendriks RJJ, Smulders MJM, Groenendael JM, Vosman B. 2000. Genetic diversity and the survival of populations. *Plant Biol*. **2**(4):379–395.

Bortoluzzi C, Bosse M, Derks MFL, Crooijmans RPMA, Groenen MAM, Megens H-J. 2020. The type of bottleneck matters. Insights into the deleterious variation landscape of small managed populations. *Evol Appl*. **13**(2):330–341.

Browning CL, Rolland RM, Kraus SD. 2010. Estimated calf and perinatal mortality in western North Atlantic right whales (*Eubalaena glacialis*). *Marine Mammal Sci*. **26**(3):648–662.

Brüniche-Olsen A, Kellner KF, Anderson CJ, DeWoody JA. 2018. Runs of homozygosity have utility in mammalian conservation and evolutionary studies. *Conserv Genet*. **19**(6):1295–1307.

Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK. Rising to the challenge of larger and richer datasets. *GigaScience* **4**:7.

Charlesworth D, Willis JH. 2009. The genetics of inbreeding depression. *Nat Rev Genet*. **10**(11):783–796.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp. An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)* **34**(17): i884–i890.

Cingolani P, Platts A, Le Wang L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **6**(2):80–92.

Coates DJ, Byrne M, Moritz C. 2018. Genetic diversity and conservation units. Dealing with the species-population continuum in the age of genomics. *Front Ecol Evol*. **6**:4045.

Cooke JG. 2018. *Balaenoptera physalus*. The IUCN Red List of Threatened Species 2018, e.T2478A50349982.

Cooke JG. 2020. *Eubalaena glacialis*. The IUCN Red List of Threatened Species 2020, e.T41712A178589687.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, *et al.* 2011. The variant call format and VCFtools. *Bioinformatics (Oxford, England)* **27**(15):2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* 2021. Twelve years of SAMtools and BCFtools. *GigaScience* **10**(2): giab008.

de Jong MJ, de Jong JF, Hoelzel AR, Janke A. 2021. SambaR. An R package for fast, easy and reproducible population-genetic analyses of biallelic SNP data sets. *Mol Ecol Resour*. **21**(4):1369–1379.

Edwards EF, Hall C, Moore TJ, Sheredy C, Redfern JV. 2015. Global distribution of fin whales *Balaenoptera physalus* in the post-whaling era (1980–2012). *Mammal Rev*. **45**(4):197–214.

Fagan WF, Holmes EE. 2006. Quantifying the extinction vortex. *Ecol Lett*. **9**(1):51–60.

Foote AD, Hooper R, Alexander A, Baird RW, Baker CS, Ballance L, Barlow J, Brownlow A, Collins T, Constantine R, *et al.* 2021. Runs of homozygosity in killer whale genomes provide a global record of demographic histories. *Mol Ecol*. **30**(23):6162–6177.

Frankham R. 2005. Genetics and extinction. *Biol Conserv* **126**(2): 131–140.

Frankham R. 2015. Genetic rescue of small inbred populations. Meta-analysis reveals large and consistent benefits of gene flow. *Mol Ecol*. **24**(11):2610–2618.

Frasier TR, Gillett RM, Hamilton PK, Brown MW, Kraus SD, White BN. 2013. Postcopulatory selection for dissimilar gametes maintains heterozygosity in the endangered North Atlantic right whale. *Ecol Evol*. **3**(10):3483–3494.

Frichot E, François O, O'Meara B. 2015. LEA. An R package for landscape and ecological association studies. *Methods Ecol Evol*. **6**(8): 925–929.

Gurgul A, Szmatoła T, Topolski P, Jasielczuk I, Żukowski K, Bugno-Poniewierska M. 2016. The use of runs of homozygosity for estimation of recent inbreeding in Holstein cattle. *J Appl Genet*. **57**(4):527–530.

Haller BC, Messer PW. 2019. SLiM 3. Forward genetic simulations beyond the Wright-Fisher model. *Mol Biol Evol*. **36**(3):632–637.

Hansen RG, Boye TK, Larsen RS, Nielsen NH, Tervo O, Nielsen RD, Rasmussen MH, Sinding MHS, Heide-Jørgensen MP. 2019. Abundance of whales in West and East Greenland in summer 2015. *NAMMCOSP* **11**.

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

Jackman SD, Coombe L, Chu J, Warren RL, Vandervalk BP, Yeo S, Xue Z, Mohamadi H, Bohlmann J, Jones SJM, *et al.* 2018. Tigmint. Correcting assembly errors using linked reads from large molecules. *BMC Bioinformatics* **19**(1):393.

Johnson JA, Tingay RE, Culver M, Hailer F, Clarke ML, Mindell DP. 2009. Long-term survival despite low genetic diversity in the critically endangered Madagascar fish-eagle. *Mol Ecol*. **18**(1): 54–63.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, *et al.* 2014. InterProScan 5. Genome-scale protein function classification. *Bioinformatics (Oxford, England)* **30**(9):1236–1240.

Jossey S, Haddrath O, Loureiro L, Lim B, Miller J, Lok S, Scherer S, Goksoyr A, Lille-Langøy R, Kovacs K, *et al.* 2021. Blue whale (*Balaenoptera musculus* musculus) genome. Population structure and history in the North Atlantic. *Authorea Preprints*.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res*. **110**(1–4):462–467.

Kardos M, Luikart G, Allendorf FW. 2015. Measuring individual inbreeding in the age of genomics. Marker-based measures are better than pedigrees. *Heredity* **115**(1):63–72.

Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PI, *et al.* 2015. Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep*. **10**(1):112–122.

Kimura M, Maruyama T, Crow JF. 1963. The mutation load in small populations. *Genetics* **48**:1303–1312.

Kraus SD, Brown MW, Caswell H, Clark CW, Fujiwara M, Hamilton PK, Kenney RD, Knowlton AR, Landry S, Mayo CA, *et al.* 2005. Ecology. North Atlantic right whales in crisis. *Science (New York, N.Y.)* **309**(5734):561–562.

Leimu R, Mutikainen PIA, Koricheva J, Fischer M. 2006. How general are positive relationships between plant population size, fitness and genetic variation? *J Ecol*. **94**(5):942–952.

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

Liu X, Fu Y-X. 2020. Stairway Plot 2. Demographic history inference with folded SNP frequency spectra. *Genome Biol*. **21**(1):280.

Lockyer C, Waters T. 1986. Weights and anatomical measurements of Northeastern Atlantic Fin (Balaenoptera physalus, Linneaus) and Sei (B. borealis Lesson) whales. *Marine Mammal Sci*. **2**(3): 169–185.

Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucl Acids Res*. **33**(20):6494–6506.

Lydersen C, Vacquié-Garcia J, Heide-Jørgensen MP, Øien N, Guinet C, Kovacs KM. 2020. Autumn movements of fin whales (*Balaenoptera physalus*) from Svalbard, Norway, revealed by satellite tracking. *Sci Rep*. **10**(1):16966.

McQuillan R, Eklund N, Pirastu N, Kuningas M, McEvoy BP, Esko T, Corre T, Davies G, Kaakinen M, Lyytikäinen L-P, *et al.* 2012. Evidence of inbreeding depression on human height. *PLoS Genet*. **8**(7):e1002655.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. 2018. Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics (Oxford, England)* **34**(13):i142–i150.

Nei M, Maruyama T, Chakraborty R. 1975. The bottleneck effect and genetic variability in populations. *Evolution* **29**(1):1.

Ochoa A, Gibbs HL. 2021. Genomic signatures of inbreeding and mutation load in a threatened rattlesnake. *Mol Ecol*. **30**(21): 5454–5469.

Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* **246**(5428):96–98.

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2. Advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics (Oxford, England)* **32**(2):292–294.

Pampoulie C, Gíslason D, Ólafsdóttir G, Chosson V, Halldórsson SD, Mariani S, Elvarsson BÞ, Rasmussen MH, Iversen MR, Daníelsdóttir AK, *et al.* 2021. Evidence of unidirectional hybridization and second-generation adult hybrid between the two largest animals on Earth, the fin and blue whales. *Evol Appl*. **14**(2): 314–321.

Paradis E, Schliep K. 2019. ape 5.0. An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)* **35**(3):526–528.

Pike DG, Gunnlaugsson T, Mikkelsen B, Halldórsson SD, Víkingsson G, Acquarone M, Desportes G. 2019. Estimates of the Abundance of Cetaceans in the Central North Atlantic from the T-NASS Icelandic and Faroese Ship Surveys Conducted in 2007. *NAMMCOSP* **11**.

Reed DH, Frankham R. 2003. Correlation between fitness and genetic diversity. *Conserv Biol*. **17**(1):230–237.

NAMMCO. 2005. Report of the Joint Meeting of the NAMMCO Working Group on North Atlantic fin whales and the IWC Scientific Committee. Nammco Annual Report. NAMMCO, Tromsø, Norway(409-442).

Robinson JA, Brown C, Kim BY, Lohmueller KE, Wayne RK. 2018. Purging of strongly deleterious mutations explains long-term persistence and absence of inbreeding depression in island foxes. *Current biology: CB* **28**(21):3487–3494.e4.

Sambrook J, Russell DW. 2006. Isolation of high-molecular-weight DNA from Mammalian cells using formamide. *CSH Protoc*. **2006**:1.

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2. Rapid adapter trimming, identification, and read merging. *BMC Res Notes* **9**:88.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO. Assessing genome assembly and annotation completeness. *Methods Mol Biol.*. (Clifton, N.J.) **1962**:227–245.

Silva MA, Prieto R, Jonsen I, Baumgartner MF, Santos RS. 2013. North Atlantic blue and fin whales suspend their spring migration to forage in middle latitudes. Building up energy reserves for the journey? *PloS one* **8**(10):e76507.

Smith G. 1984. The International Whaling Commission. An analysis of the past and reflections on the future. *Nat Resour Lawyer* **16**: 543–567.

Spielman D, Brook BW, Briscoe DA, Frankham R. 2004a. Does inbreeding and loss of genetic diversity decrease disease resistance? *Conserv Genet.* **5**(4):439–448.

Spielman D, Brook BW, Frankham R. 2004b. Most species are not driven to extinction before genetic factors impact them. *Proc Natl Acad Sci U S A* **101**(42):15261–15264.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS. *Ab initio* prediction of alternative transcripts. *Nucl Acids Res.* **34**(Web Server issue):W435–W439.

Tanaka Y. 2000. Extinction of populations by inbreeding depression under stochastic environments. *Population Ecol.* **42**(1):55–62.

Taylor BL, Chivers S, Larese J, Perrin W. 2007. Generation length and percent mature estimates for IUCN assessments of cetaceans. Southwest Fisheries Science Center Administrative Report. (LJ-07-01).

Teixeira JC, Huber CD. 2021. The inflated significance of neutral genetic diversity in conservation genetics. *Proc Natl Acad Sci U S A* **118**(10):e2015096118.

Tollis M, Robbins J, Webb AE, Kuderna LFK, Caulin AF, Garcia JD, Bèrubè M, Pourmand N, Marques-Bonet T, O'Connell MJ, *et al.* 2019. Return to the Sea, Get Huge, Beat Cancer. An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). *Mol Biol Evol.* **36**(8):1746–1763.

Tønnessen JN, Johnsen AO. 1982. *The history of modern whaling*. Berkeley: University of California Press.

van der Valk T, de Manuel M, Marques-Bonet T, Guschanski K. 2019a. Estimates of genetic load in small populations suggest extensive purging of deleterious alleles. *bioRxiv*. 696831.

van der Valk T, Díez-Del-Molino D, Marques-Bonet T, Guschanski K, Dalén L. 2019b. Historical genomes reveal the genomic consequences of recent population decline in Eastern Gorillas. *Curr Biol: CB* **29**(1):165–170.e6.

von Seth J, Dussex N, Díez-Del-Molino D, van der Valk T, Kutschera VE, Kierczak M, Steiner CC, Liu S, Gilbert MTP, *et al.* 2021. Genomic insights into the conservation status of the world's last remaining Sumatran rhinoceros populations. *Nat Commun.* **12**(1):2393.

Webb A, Knoblauch J, Sabankar N, Kallur AS, Hey J, Sethuraman A. 2021. The Pop-Gen Pipeline Platform. A software platform for population genomic analyses. *Mol Biol Evol.* **38**(8): 3478–3485.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. *Genome Res.* **27**(5): 757–767.

Westbury MV, Hartmann S, Barlow A, Wiesel I, Leo V, Welch R, Parker DM, Sicks F, Ludwig A, Dalén L, *et al.* 2018. Extended and continuous decline in effective population size results in low genomic diversity in the World's Rarest Hyena Species, the Brown Hyena. *Mol Biol Evol.* **35**(5):1225–1237.

Westbury MV, Petersen B, Garde E, Heide-Jørgensen MP, Lorenzen ED. 2019. Narwhal genome reveals long-term low genetic diversity despite current large abundance size. *iScience* **15**: 592–599.

Yeo S, Coombe L, Warren RL, Chu J, Birol I. 2018. ARCS. Scaffolding genome drafts with linked reads. *Bioinformatics (Oxford, England)* **34**(5):725–731.

Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, Oh H-M, Lee J-H, Yang EC, Kwon KK, *et al.* 2014. Minke whale genome and aquatic adaptation in cetaceans. *Nat Genet.* **46**(1):88–92.

Supplementary Materials for

Genomic impact of whaling in North Atlantic fin whales.

Magnus Wolf, Menno de Jong, Sverrir Daníel Halldórsson, Úlfur Árnason, Axel Janke

*Corresponding author: Magnus Wolf; Email: Magnus.Wolf@senckenberg.de

This PDF file includes:

Figs. S1 to S8

Tables S1 to S4

References used in this document are all included in the main manuscript.

Supplementary Figures

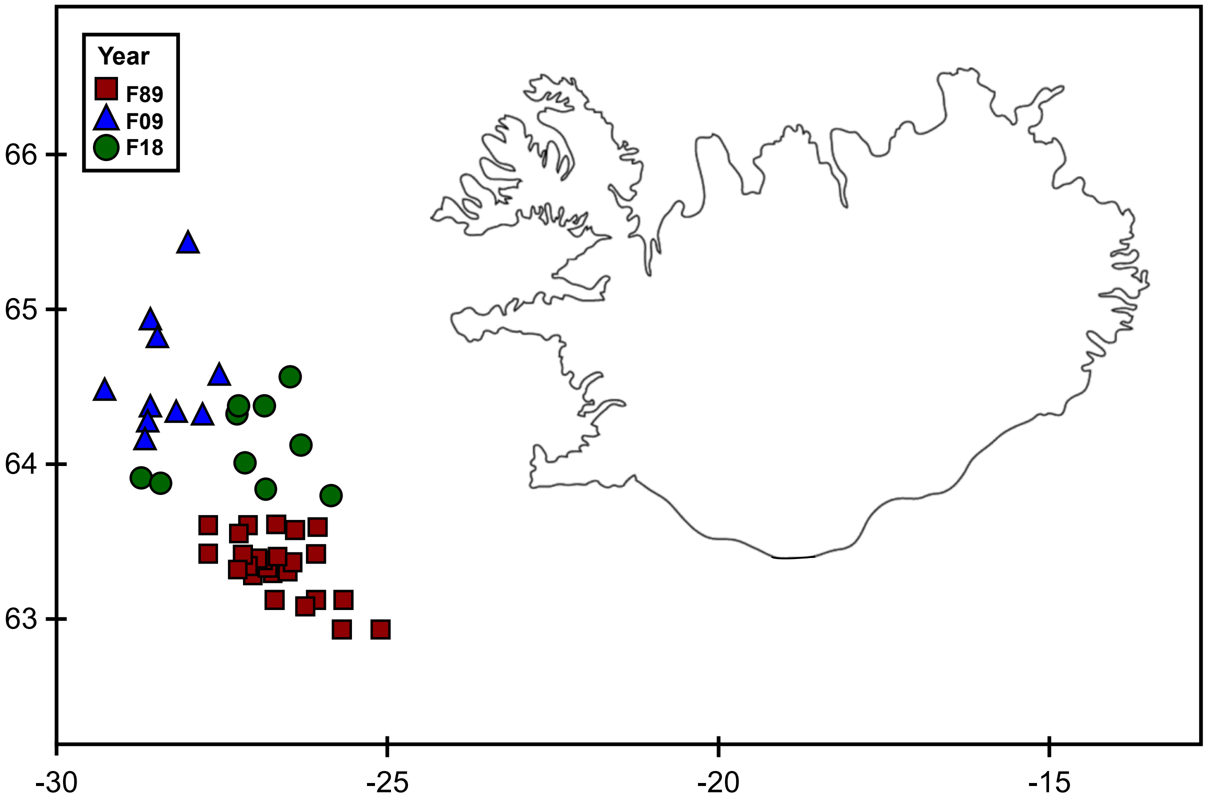

Fig. S1 Geographical distribution of sampled whales presented in this study. F and the following number in the upper left box denotes the sampling year (1989: red, 2009: blue, 2018: green) of a fin whale (F). Numbers on the coordinates denote the latitude and longitude (WGS84), shown in Table S1.

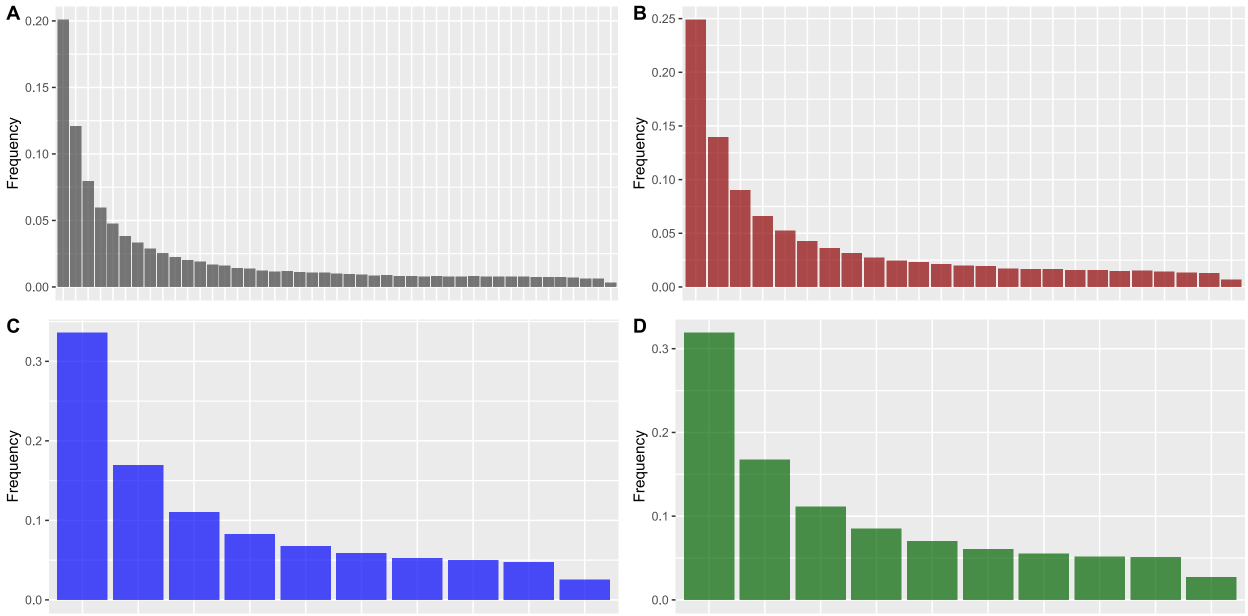

Fig. S2 Folded site frequency spectrum (fSFS) for all fin whale samples combined (A, grey) and the three fin whale cohorts 1989 (B, red), 2009 (C, blue) and 2018 (D, green) separately. Bars indicate the proportion of polymorphic sites for a given number of minor allele copies. fSFS were used for downstream analyses like the demographic models estimated by STAIRWAY PLOT v2 (Liu and Fu 2020).

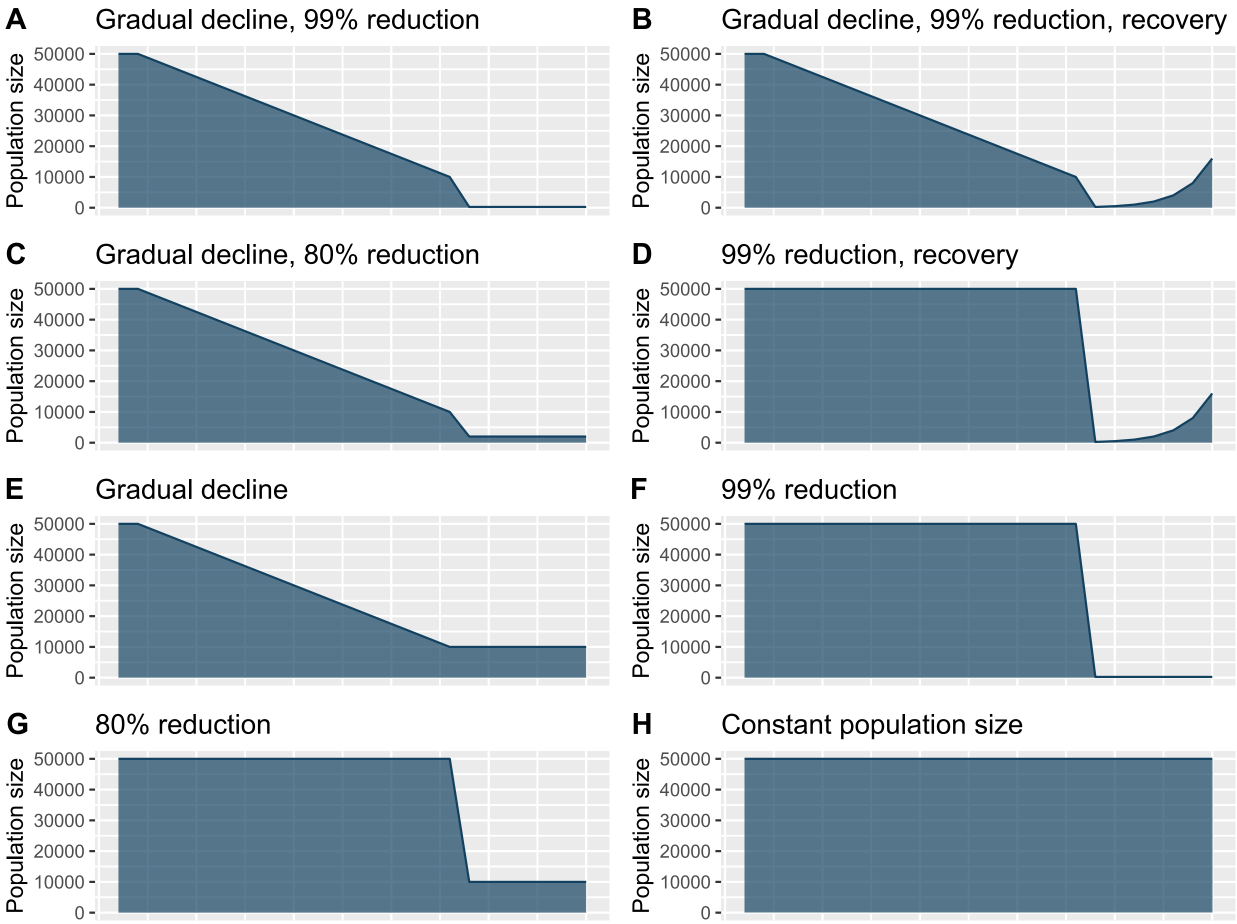

Fig. S3 Demographic scenarios simulated with forward-in-time Wright-Fischer model simulations using SLiM (Haller et al. 2019), arranged in order of model performance (Table S4). Model performance was estimated as the log-likelihood of observing the empirical fSFS given the simulated fSFS. Demographic scenarios involving migration between a bottlenecked and a non-bottlenecked population performed poorly and have been omitted here.

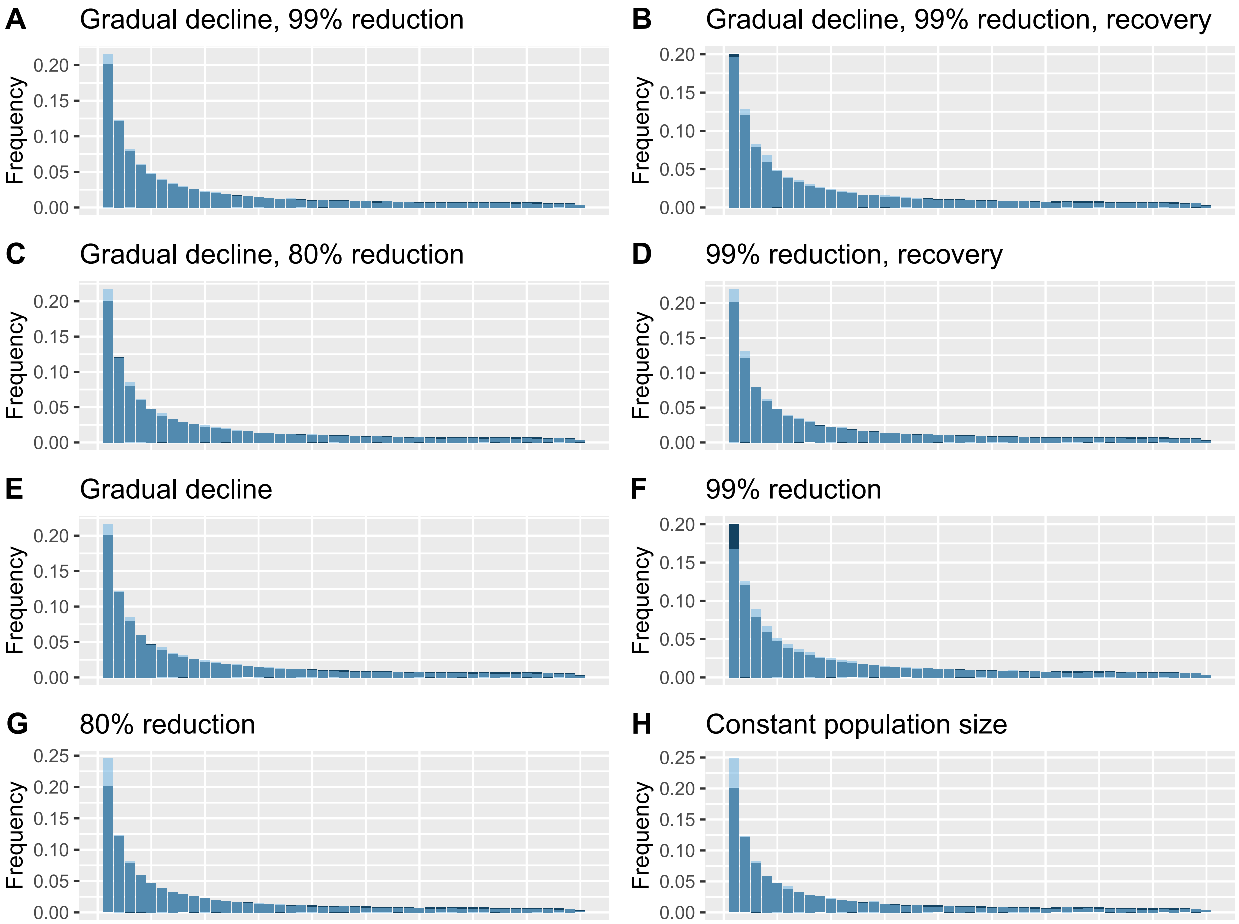

Fig. S4 Comparison between the empirical fSFS (Fig. S2) and the fSFS predicted by forward-in-time Wright-Fischer model simulations using SLiM (Haller et al. 2019), arranged in order of model performance (i.e., log-likelihood score, Table S4). Darkblue: empirical fSFS; lightblue predicted fSFS.

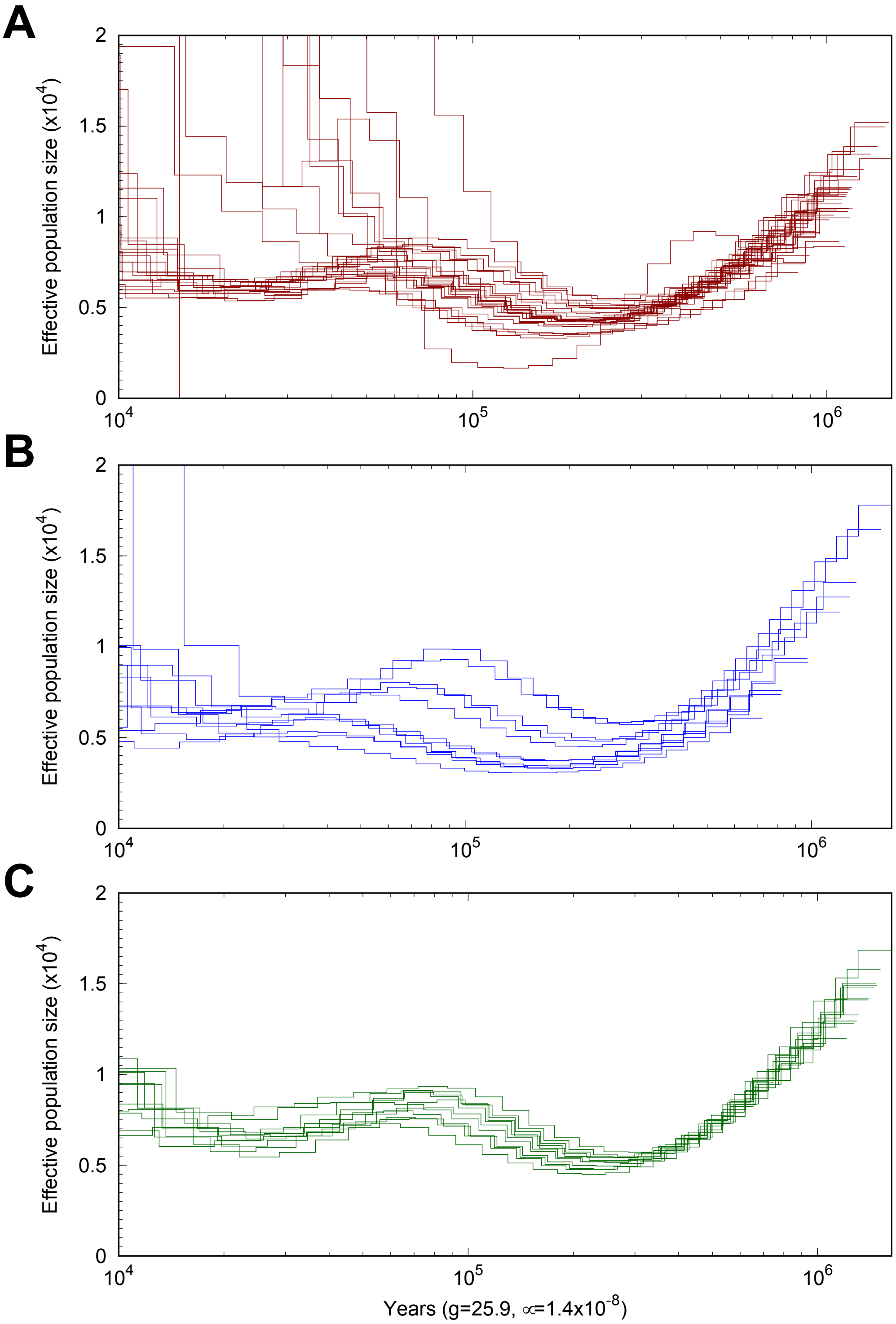

Fig. S5 Changes in N_e over the last 1 Mya to 10 kya compared among the fin whale cohorts (color code matches main manuscript) using a PSMC analysis (Li and Durbin 2011). For the three cohorts, the trajectory follows the results shown in Árnason et al. (2018), showing a rather stable, slightly fluctuating N_e over time. Plots were scaled using a mutation rate of 1.54 x 10^{-9} per site per generation and a generation time of 25.9 years.

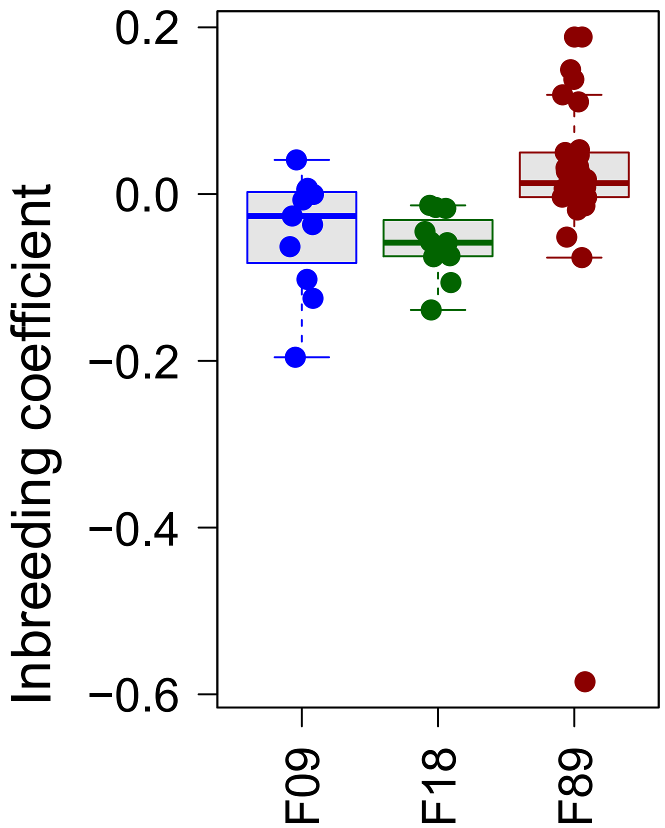

Fig. S6 Distribution of inbreeding coefficients F_H (Kardos et al. 2015) for the three fin whale cohorts sampled in the years 1989 (red), 2009 (blue) and 2018 (green). The mean value was slightly negative for all cohorts. However, in the 2009 and 2018 cohort, there were more negative inbreeding factors suggesting slight excess in heterozygous genotypes compared to the expected value zero.

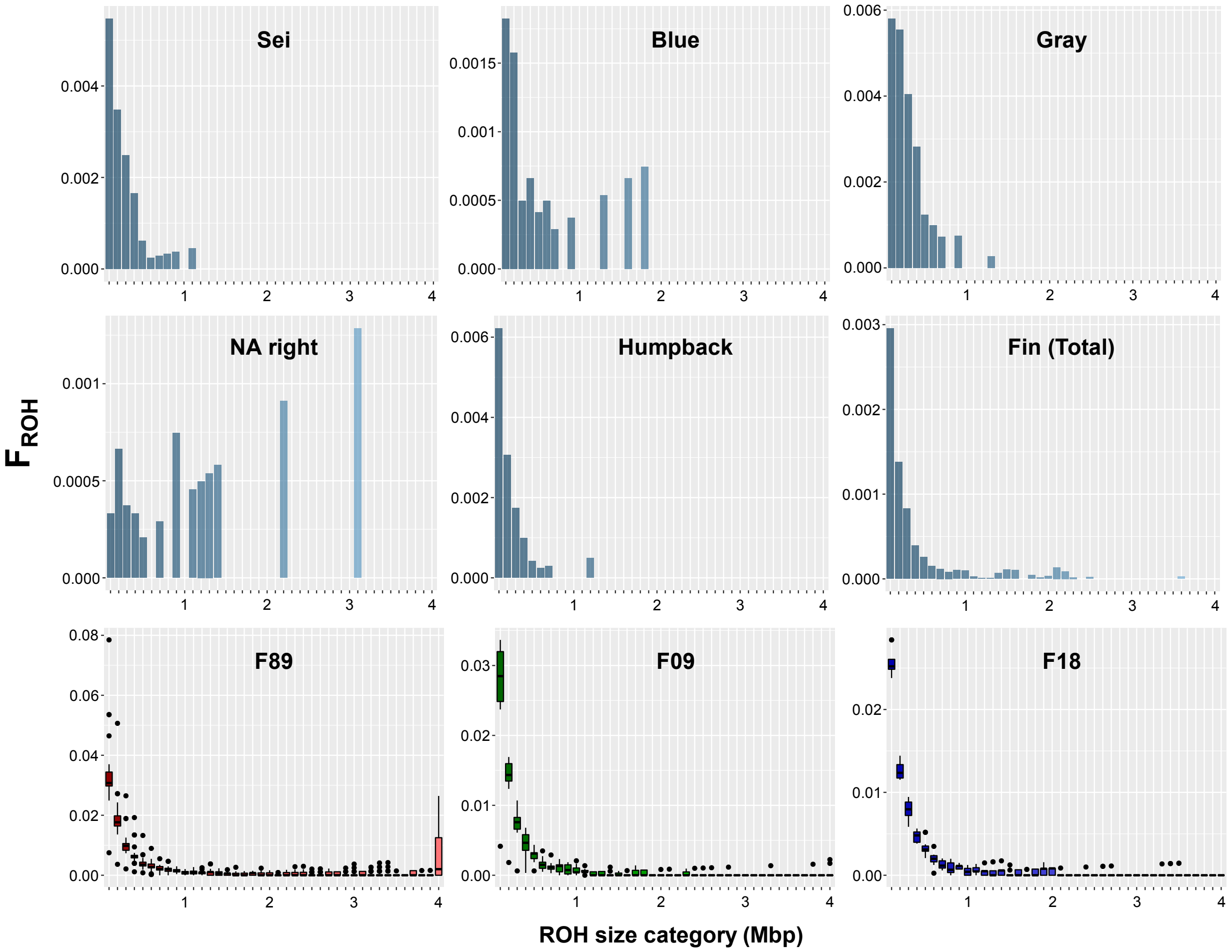

Fig. S7 Inbreeding factors (F_{ROH}) based on the genome coverage of run of homozygosity (ROH) between different minimal lengths cutoffs of ROH: 100 kbp to over 4 Mbp (x-axis in 100 kbp steps).

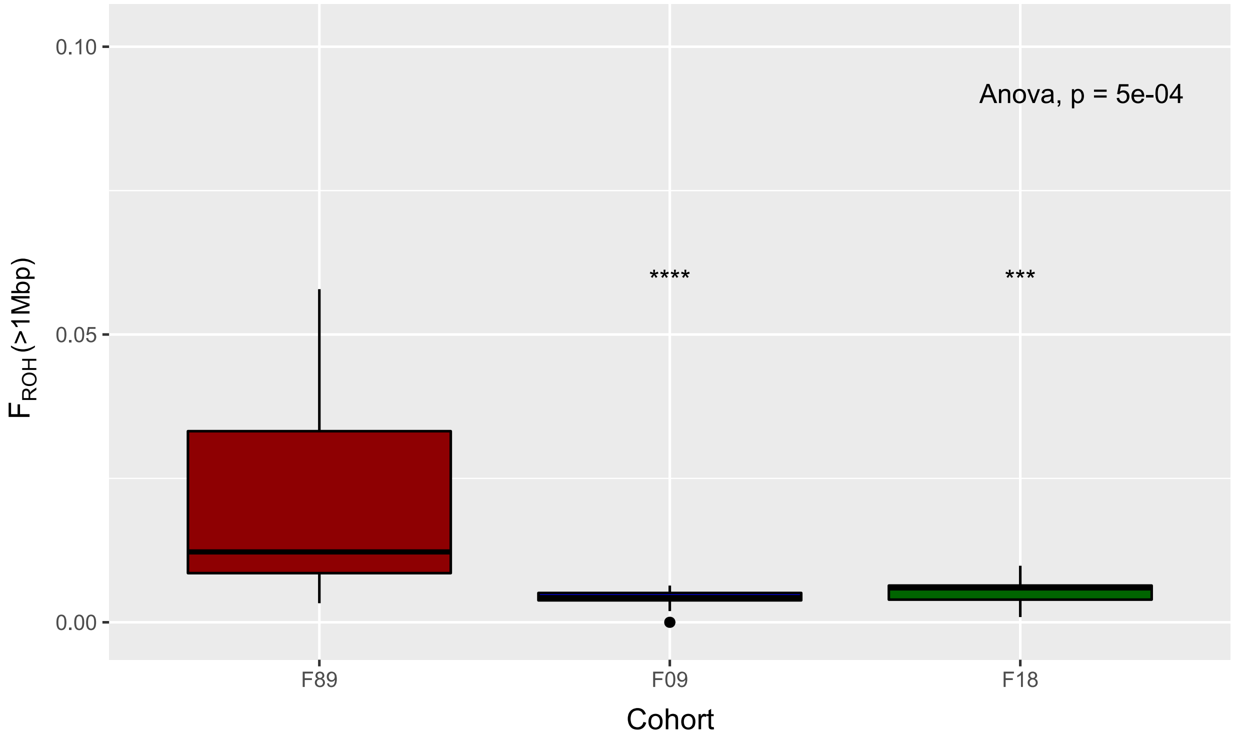

Fig. S8 ANOVA significance test comparing inbreeding coefficients (F_{ROH}) of three fin whale cohorts (1989: red, 2009: blue, 2018: green) based on runs of homozygosity longer than 1 Mbp. The total proportion of the reference genome covered by those ROH varied significantly between 1989 and both other cohorts (2009: *p=0.003*; 2018: *p=0.006*), respectively. Between the 2009 and 2018 cohort, no significant differences were found. Following these results, it can be assumed that some individuals of the 1989 cohort experienced inbreeding in the more recent past.

Supplementary Tables

Table S1 General information of sampled fin whales that were sequenced and analyzed in this study. Included are the used sample-ID, sex as well as the specific capture location and date**.**

Sample-ID	Sex	Latitude	Longitude	Date
F89001	f	63.6	26.2	19. Jun. 89
F89002	f	63.6	27.8	21. Jun. 89
F89003	m	63.4	27.8	21. Jun. 89
F89004	f	63.1	26.8	21. Jun. 89
F89005	f	63.4	27.2	21. Jun. 89
F89006	f	63.6	26.8	22. Jun. 89
F89007	f	63.6	26.8	23. Jun. 89
F89008	f	63.6	26.5	23. Jun. 89
F89009	f	63.4	26.8	23. Jun. 89
F89010	f	63.4	26.8	24. Jun. 89
F89011	f	63.4	26.8	24. Jun. 89
F89012	f	63.4	27.2	24. Jun. 89
F89013	m	63.4	26.8	24. Jun. 89
F89014	f	63.4	26.8	25. Jun. 89
F89015	m	63.4	26.8	25. Jun. 89
F89016	m	63.4	26.8	25. Jun. 89
F89017	f	63.1	26.8	26. Jun. 89
F89018	f	63.4	26.2	27. Jun. 89
F89019	m	63.4	26.2	27. Jun. 89
F89020	f	63.6	27.2	27. Jun. 89
F89030	m	62.9	25.2	1. Jul. 89
F89031	m	62.9	25.8	1. Jul. 89
F89032	m	62.9	25.8	1. Jul. 89
F89033	f	63.1	25.8	2. Jul. 89
F89035	f	63.1	26.8	3. Jul. 89
F89036	f	63.4	26.8	3. Jul. 89
F89037	f	63.1	26.2	3. Jul. 89
F89038	m	63.1	26.2	3. Jul. 89
F09019	f	65.4	28.1	03. Jul 09
F09020	m	64.6	27.6	04. Jul 09
F09021	f	64.3	28.3	05. Jul 09
F09022	m	64.2	28.7	05. Jul 09
F09023	m	64.3	28.7	06. Jul 09
F09024	f	64.2	28.7	06. Jul 09
F09025	m	64.8	28.6	08. Jul 09
F09026	f	64.9	28.6	08. Jul 09
F09027	m	64.3	27.9	09. Jul 09
F09028	m	64.5	29.3	10. Jul 09
F09029	f	64.5	29.4	10. Jul 09
F18003	m	64.4	26.9	24.Jun 18
F18004	f	64.1	26.4	24.Jun 18
F18006	f	64.4	27.4	28. Jun 18
F18007	m	64.3	27.4	28. Jun 18
F18008	f	64.6	26.6	29. Jun 18
F18009	m	63.9	28.5	30. Jun 18
F18010	f	63.9	28.8	30. Jun 18
F18011	m	64.0	27.2	03. Jul 18
F18013	f	63.8	26.0	03. Jul 18
F18014	m	63.8	26.0	03. Jul 18
F18015	m	63.8	26.9	05. Jul 18

f, female; m, male.

Table S2 Summary statistics, BUSCO completeness analyses and annotation statistics for the fin whale reference genome.

Assembly statistics	
No. contigs	**13,639**
No. contigs (>50 kbp)	**305**
L50	**27**
L75	**58**
N50 (bp)	**24,939,677**
N75 (bp)	**14,958,115**
Max.contig length (bp)	**91,471,248**
Total length (bp)	**2,412,335,827**
GC (%)	**40.81**
No. of N's per 100 kbp	**1,702.22**
BUSCO completeness	
BUSCO (*Cetartiodactyla*)	**C:83.4%[S:82.1%,D:1.3%],** **F:4.1%,M:12.5%,** **n:13335**
BUSCO (*Laurasiatheria*)	**C:90.5%[S:89.0%,D:1.5%]** **F:3.0%,M:6.5%** **n:12234**
BUSCO (*Mammalia*)	**C:91.2%[S:89.9%,D:1.3%]** **F:2.5%,M:6.3%** **n:9226**
Annotation statistics	
Total interspersed repeats (bp)	**1,008,206,033 (41.79 %)**
Number of transcripts	**17,307**
Functional annotated genes	**17,152 (99.1%)**

BUSCO: Benchmarking Universal Single Copy Orthologs (Seppey et al. 2019); C, complete; S, single copy; D, duplicated; F, fragmented; M, missing.

Table S3 Repeat content of the fin whale assembly compiled for this study. Repeats were collected by REPEATMASKER v4.1 (www.repeatmasker.org) after modelling and identifying them with REPEATMODELER v2 (www.repeatmasker.org) using the Cetartiodactyla database from REPBASE (Jurka et al. 2005).

	Number of elements	Length occupied (bp)	Percentage of sequence
Retroelements	2286384	918589536	38.08
SINEs:	666521	121878092	5.05
Penelope	65	13180	0.00
LINEs:	1235730	649137675	26.91
CRE/SLACS	0	0	0.00
L2/CR1/Rex	254400	65703824	2.72
R1/LOA/Jockey	0	0	0.00
R2/R4/NeSL	402	105485	0.00
RTE/Bov-B	10799	3251621	0.13
L1/CIN4	969766	579940256	24.04
LTR elements:	384133	147573769	6.12
BEL/Pao	0	0	0.00
Ty1/Copia	1086	2740340	0.11
Gypsy/DIRS1	14835	4166645	0.17
Retroviral	357392	137801181	5.71
DNA transposons:	384998	84056461	3.48
hobo-Activator	274224	55671684	2.31
Tc1-IS630-Pogo	97076	26618088	1.10
En-Spm	0	0	0.00
MuDR-IS905	0	0	0.00
PiggyBac	631	216795	0.01
Tourist/Harbinger	334	60928	0.00
Other (Mirage,Pelement,Transib)	0	0	0.00
Rolling-circles	1597	407845	0.02
Unclassified:	29118	5560036	0.23
Total interspersed repeats:		1008206033	41.79
Small RNA:	292610	66915824	2.77
Satellites:	3649	1767591	0.07
Simple repeats:	12507	2299460	0.10
Low complexity:	0	0	0.00

Table S4 Log-likelihood scores showing the probability of observing the empirical fSFS given the predicted fSFS for each forward-in-time Wright-Fischer model simulation using SLiM (Haller et al. 2019). Log-likelihood scores have been calculated using the R base function ‘dmultinom’.

Scenario	Previous Demography	Whaling Decimation	Following Demography	Migration	Log-Likelihood
1	Gradual decline	99%	-	-	-1.207,963
2	Gradual decline	99%	Recovery	-	-1.327,108
3	Gradual decline	80%	-	-	-1.327,108
4	-	99%	Recovery	-	-1.350,702
5	Gradual decline	-	-	-	-1.474,814
6	-	99%	-	-	-1.773,815
7	-	80%	-	-	-2.353,175
8	-	-	-	-	-3.039,545
9	Gradual decline	99%	-	slow Migration	-10.530,82
10	-	99%	-	fast Migration	-10.717,62
11	Gradual decline	99%	-	fast Migration	-11.776,07
12	-	80%	-	slow Migration	-11.969,25
13	Gradual decline	99%	Recovery	fast Migration	-13.797,7
14	Gradual decline	80%	-	slow Migration	-14.187,21
15	Gradual decline	-	-	slow Migration	-14.914,1
16	Gradual decline	80%	-	fast Migration	-15.270,46
17	Gradual decline	-	-	fast Migration	-15.565,79
18	-	80%	-	fast Migration	-17.072,05
19	-	-	-	slow Migration	-17.599,82
20	-	80%	-	slow Migration	-18.100,24
21	-	-	-	fast Migration	-18.368,42

6.2 Publication 2:

The genome of the pygmy right whale illuminates the evolution of rorquals.

Magnus Wolf, Konstantin Zapf, Deepak Kumar Gupta, Michael Hiller, Úlfur Árnason, Axel Janke

Published as a research article in: *BMC Biology* **21**, 79 (12 April 2023).

https://doi.org/10.1186/s12915-023-01579-1

Declaration of author contributions to the publication

Publication: The genome of the pygmy right whale illuminates the evolution of rorquals.

Status: published

Name of the journal: BMC Biology

Contributing Authors:

Magnus Wolf (MW), Konstantin Zapf (KZ), Deepak Kumar Gupta (DKG), Michael Hiller (MH), Úlfur Árnason (UA), Axel Janke (AJ)

Contributions of the doctoral candidate and his co-authors:

1.) Concept and design		
MW	70%	
UA	10%	
AJ	20%	
2.) Data analysis and figure compilation		
MW	80%	All phylogenomic and selection analysis
KZ	10%	Assembly construction
DKG	5%	Assembly construction
MH	5%	Whole Genome Alignment construction
3.) Interpretation of data		
MW	80%	Interpretation of all data
AJ	20%	Summary interpretation
4.) Drafting manuscript		
MW	70%	
MH	5%	
UA	5%	
AJ	20%	
5.) Other works		
UA	100%	Sample organisation

I hereby certify that the information above is correct.

____________________ ____________________

Date and place Signature doctoral candidate

____________________ ____________________

Date and place Signature supervisor

BMC Biology

RESEARCH ARTICLE

Open Access

The genome of the pygmy right whale illuminates the evolution of rorquals

Magnus Wolf[1,2*], Konstantin Zapf[1,2], Deepak Kumar Gupta[3], Michael Hiller[1,3,4], Úlfur Árnason[5,6] and Axel Janke[1,2,3]

Abstract

Background Baleen whales are a clade of gigantic and highly specialized marine mammals. Their genomes have been used to investigate their complex evolutionary history and to decipher the molecular mechanisms that allowed them to reach these dimensions. However, many unanswered questions remain, especially about the early radiation of rorquals and how cancer resistance interplays with their huge number of cells. The pygmy right whale is the smallest and most elusive among the baleen whales. It reaches only a fraction of the body length compared to its relatives and it is the only living member of an otherwise extinct family. This placement makes the pygmy right whale genome an interesting target to update the complex phylogenetic past of baleen whales, because it splits up an otherwise long branch that leads to the radiation of rorquals. Apart from that, genomic data of this species might help to investigate cancer resistance in large whales, since these mechanisms are not as important for the pygmy right whale as in other giant rorquals and right whales.

Results Here, we present a first de novo genome of the species and test its potential in phylogenomics and cancer research. To do so, we constructed a multi-species coalescent tree from fragments of a whole-genome alignment and quantified the amount of introgression in the early evolution of rorquals. Furthermore, a genome-wide comparison of selection rates between large and small-bodied baleen whales revealed a small set of conserved candidate genes with potential connections to cancer resistance.

Conclusions Our results suggest that the evolution of rorquals is best described as a hard polytomy with a rapid radiation and high levels of introgression. The lack of shared positive selected genes between different large-bodied whale species supports a previously proposed convergent evolution of gigantism and hence cancer resistance in baleen whales.

Keywords Pygmy right whale, *Caperea marginata*, Whole-genome sequencing, Phylogenomics, Rorquals, Positive selection, Cancer resistance, Peto's paradox

*Correspondence:
Magnus Wolf
Magnus.Wolf@senckenberg.de
Full list of author information is available at the end of the article

Teaser

The genome of the smallest baleen whale was used to update the rorqual phylogeny and to identify genes related to cancer resistance.

Background

Biology of *Caperea marginata*

Baleen whales (*Mysticeti*) are the largest animals on earth, reaching up to 30 m in length and a weight of 150 metric tons. These iconic animals have received considerable public and scientific interest in the past. The pygmy right whale (*Caperea marginata*, Gray 1846) is the smallest species among the baleen whales, with records ranging between 5 and 6.5 m in length and weighing 3t to 3.5t [1]. They have a circumpolar distribution around the southern hemisphere, although crossing equatorial regions and hence a wider distribution may be possible [2]. The biology of this species is still poorly understood, not only because the number of sightings is very limited, but also because of possible confusions with minke whales (*Balaenoptera acutorostrata* and *B. bonaerensis*) [1]. They are believed to be slow swimming filter feeders given the morphology of their feeding apparatus that is similar to that of the right whales (*Balaenidae*) [3, 4] and it is assumed that they are not deep divers because their heart and lungs are, compared to such whales, relatively small [5]. There is no information available about their abundance, but they were not targeted by whalers in the past [1]. The species is listed as “least concern” on the IUCN red list due to the lack of information [6].

Phylogeny of *Caperea marginata*

Multiple features of the skull and skeletal morphology of the pygmy right whale are unique within the baleen whales [7], leading to a complex history of taxonomic re-assignments and re-naming of the species [8]. Many morphological studies suggested that the species is an early diverging member of the right whales [9–11], hence its common name: pygmy right whale. However, molecular studies have consistently reported a closer relationship to the rorquals, *Balaenopteridae* [12–17]. In recent years, the species was allocated closer to the rorquals and placed in the otherwise extinct family of *Cetotheriidae* and the subfamily of *Neobalaenidae* based on a combination of extensive comparisons to fossil records and on molecular data [3]. This placement was later supported by a phylogenomic study that included nearly all extant *Cetacea* species [18].

Phylogeny problems of rorquals

While placing the pygmy right whale in the *Cetotheriidae* and hence the *Neobalaenidae* seems resolved, placing and ordering groups within the *Balaenopteridae*, rorquals, remains challenging. It is assumed that the rorquals experienced a rapid radiation in combination with incomplete lineage sorting (ILS) and introgression after diverging from a common ancestor 10 to 25 million years ago [14, 18–20]. Especially the phylogenetic position of the gray whale (*Balaenoptera robustus*, formerly *Eschrichtius robustus*, see Årnason et al. (2018) [20]) remains uncertain because even recent phylogenomic analyses addressing this clade resulted in short branches with low support despite their plethora of molecular data [18, 20]. However, current studies were either limited by taxon sampling due to the lack of whole-genome sequences [20] or might have been hampered by the limited amount of evolutionary information per short protein-coding sequences [18]. Revisiting the problematic phylogeny of rorquals with an increased taxon-sampling and long whole-genome alignment (WGA) fragments rather than short coding sequences is expected to increase the resolution of this rapidly diverged species complex. The genome of the pygmy right whale will most likely improve the resolution of rorqual evolution because of its placement at the base of the rorqual divergence and its addition to phylogenomic analyses will split up an otherwise long branch separating rorquals from right whales.

Peto’s paradox and cancer research

Baleen whales have also received substantial interest in research regarding cancer resistance because of their relatively normal cancer mortality despite their large number of cells and relatively long life-expectancy known as the “Peto’s” paradox [21–24]. In previous attempts, the identification of related oncogenes and tumor suppressor genes (TSG) was often based on a genome-wide comparison of selection rates or gene copy numbers between a large species and a smaller relative [22, 24, 25]. Although these pairwise comparisons resulted in numerous candidate genes that may be responsible for the resistance to cancer in baleen whales, their identification has remained vague because previous studies were restricted to a single species pair, given that only the minke whale genome [26] was available as a small-bodied reference. Alternative approaches exist, such as codon-based models that estimate selective pressure along evolutionary branches [22, 24], but they need to be treated with caution because model misspecification and alignment errors can result in a potentially high number of artifacts [27]. Increasing the number of pairs available for selection rate comparisons could dramatically increase their precision. Therefore, adding a reference genome for the considerably smaller pygmy right whale will be a valuable addition to this kind of research and will likely reduce the risk of identifying false positive candidate genes.

Objectives

In this study, we assemble a de novo reference genome for the pygmy right whale (*Caperea marginata,* Gray 1846) and test its potential to improve the phylogenetic resolution among the rorquals and to increase the chance to identify potential cancer-related genes. Therefore, we include the new genome in a reissued phylogenomic analysis with an increased sampling of whole-genome sequences and long WGA fragments rather than short coding sequences. A set of candidate genes related to cancer resistance is compiled by comparing selection rates between several large-bodied baleen whales and small-bodied relatives. Additionally, we provide a first estimate of the genetic diversity and model the demographic history of the species to provide new insights into this elusive species.

Table 1 Summary statistics, BUSCO completeness analysis, and annotation statistics for the *C. marginate* reference genome

Assembly statistics	
No. contigs	**51,950**
No. contigs (>50 kbp)	**16,071**
L50	**9883**
N50 (bp)	**112,264**
Total length (bp)	**2,515,163,484**
GC (%)	**41.19**
No. of *N*s per 100 kb	**8920.26**
Heterozygosity (%)	**0.11**
BUSCO completeness	
BUSCO (cetartiodactyla)	**C: 63.7%[S: 61.9%, D:1.8%]** **F: 9.9%, M: 26.4%** ***n*: 13,335**
BUSCO (laurasiatheria)	**C: 66.7%[S: 64.7%, D: 2.0%]** **F: 14.3%, M: 19.0%** ***n*: 12,234**
BUSCO (mammalia)	**C: 65.6%[S: 63.6%, D: 2.0%]** **F: 14.1%, M: 20.3%** ***n*: 9226**
Annotation statistics	
Total interspersed repeats (bp)	**951,069,063 (37.81%)**
Number of transcripts	**32,808**
Functional annotated genes	**28,267 (86.1%)**

BUSCO Benchmarking Universal Single Copy Orthologs, *C* Complete, *S* Single copy, *D* Duplicated, *F* fragmented, *M* Missing

Results

Genome characteristics

The genome of the pygmy right whale was assembled to a total size of 2.5 Gbp and consists of 51,945 contigs with an N50 of 112.3 kbp and an L50 of 6438 contigs (Table 1). The GC content of the finale assembly is 41.2%, scaffolds contain 8920.3 N's per 100 kbp and the genome-wide heterozygosity is 0.11%. BUSCO completeness analyses of the three OrthoDB clades *Cetartiodactyla*, *Laurasiatheria*, and *Mammalia* yielded 63.7%, 66.7%, and 65.7% complete core genes, respectively, as well as 9.9%, 14.3%, and 14.1% fragmented genes. A de novo modeling and masking of repeats found that 37.8% of the genome is covered by interspersed repeats (Additional file 1: Table S1). Homology-based annotation yielded 33,644 potential transcripts, of which 95.7% were functionally annotated.

Phylogenomics

A multispecies coalescent (MSC) phylogenomic tree (Fig. 1A) was constructed for the entire *Mysticeti* including whole-genomes from twelve extant baleen whale species. The tree was conflated from 46,941 individual maximum likelihood (ML) trees that were each constructed from 20 kbp fragments cut from an 1.3 Gbp long whole-genome alignment (WGA) using the genome of the bottlenose dolphin (*Tursiops truncatus*) as a reference. Each fragment contains a mean of ~ 730 parsimony informative sites and the ideal fragment size was determined to be 20 kbp using an approximately unbiased test (Additional file 1: Fig. S1). All branches of the final tree are supported with local posterior probabilities of 1.0 and final branch lengths were estimated using a maximum likelihood inference based on 563 high-quality shared single-copy orthologous amino acid sequences (SCOS). The resulting tree depicts a clear separation between the three baleen whale families: *Balaenidae,*

(See figure on next page.)

Fig. 1 Phylogenomic analysis of baleen whales using whole-genome alignment fragments. **A** Phylogenomic multi-species coalescent (MSC) tree inferred from 46,941 trees that were each constructed from a 20 kbp whole-genome alignment fragment. All branches conceived 1.0 local posterior probabilities and branch lengths were added by a maximum likelihood inference using amino acid sequences of 563 high-quality single copy ortholog sequences. The pygmy right whale was placed at the base of the rorquals, and the gray whale was grouped together with the humpback whale and fin whale. **B** Quartet scores of different branches across the MSC tree. Branches 1– 7 were analyzed for the number of trees supporting one of the three possible unrooted topologies (q1– q3). Branch 4, representing the position of the gray whale, received nearly equal quartet scores for all alternative topologies. **C** Distribution of quartet scores across the first 20 Mbp of chromosome one of the reference assembly (*Tursiops truncatus*), given three different topologies of branch 4 (light to dark blue). Across the chromosome, no clear runs of shared phylogenetic signals could be identified. Pygmy right whale illustration made by Frédérique Lucas. The assembly data used to generate the results shown can be found in Additional file 1: Table S6 [28–38]

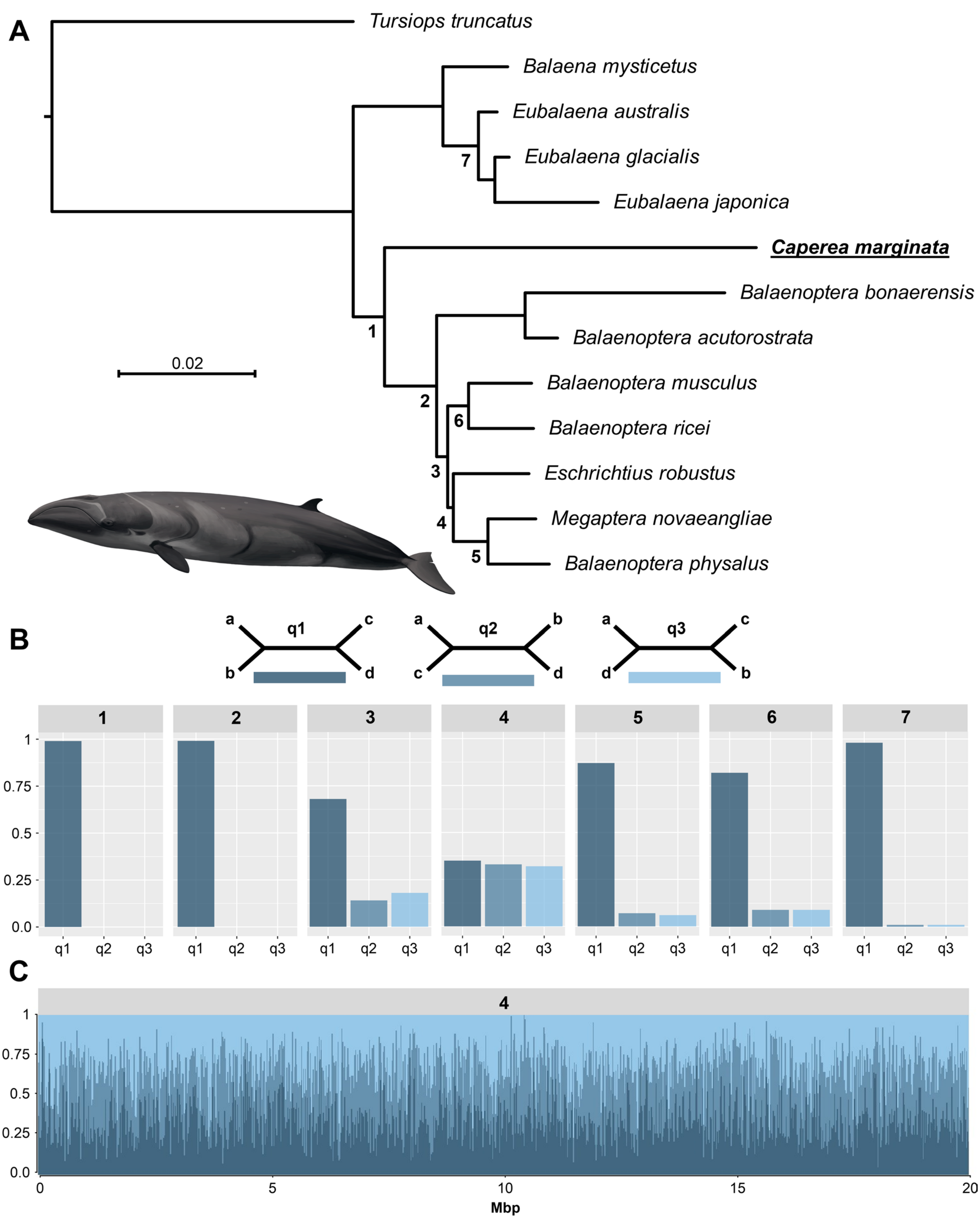

Fig. 1 (See legend on previous page.)

Balaenopteridae, and *Cetotheriidae.* Within the rorquals, the tree supports a grouping of the gray whale with the pair of fin whale and humpback whale, while a sister clade is formed by the blue whale and the rice whale.

For most branches, we found low support for alternative topologies, represented by quartet scores that describe the conflicts between the three possible topologies for an internal, unrooted branch (Fig. 1B). The placement of the gray whale received nearly identical support for all three topologies, with only a slight excess towards the topology presented in the MSC tree. To depict the distribution of these conflicts across the genome, we calculated quartet scores for every 20 kbp fragment of the largest reference chromosome 1, given the overall MSC tree. This analysis revealed an even distribution of signals between the three possible topologies (Fig. 1C) over the entire chromosome with no clear runs of shared phylogenetic signals.

The conflicting phylogenetic signals are also evident within a consensus network constructed from the entire set of 46,941 WGA fragments (Fig. 2). Using a threshold of 12% conflicting edges, we received conflicting topologies at the base of the *Balaenidae, Balaenopteridae*, and for the placement of the gray whale with the latter having the most even distribution of conflicts. Lower thresholds resulted in more complex patterns indicating additional, though less frequent, conflicting phylogenetic signals from individual WGA fragments (Additional file 1: Fig. S2).

To evaluate whether these conflicting signals originated from ILS or introgression, we quantified introgression events via branch lengths (QuIBL) as presented by Edelman et al. 2019 [39]. Based on the distribution of internal branch lengths of discordant triplet topologies, the test determines whether conflicting trees were the result from either ILS only (H_0), or ILS together with introgression (H_1). In the case of ILS only (H_0), branch lengths are expected to be exponentially distributed. Restricting ourselves to all rorquals in close evolutionary proximity to the contested gray whale, we found evidence for introgression ($\Delta BIC < -10$) in 5 out of 10 possible triplets (Additional file 1: Table S2). We found on average 33% discordant WGA fragment trees per triplet of which ~ 58% were likely the result of past introgression events. Thus, of the total number of evaluated trees, a mean of 19% (H_1) is estimated to be affected by introgression while the other 14% showed an exponential distribution of internal branch lengths and are therefore considered to be the result of ILS only (H_0). Triplets that included the gray whale as well as one representative of both possible sister clades usually showed around 66% discordant trees of which 64%, or 42% of the total data set, likely had a history of introgression. The triplet of gray whale, fin whale, and blue whale resulted in the most signals of introgression with 48.1% of all tested trees. Hence, we assume introgression to be the dominant driver for conflicting trees in our dataset, especially around the contested position of the gray whale.

Additionally, we constructed an MSC tree from 563 maximum likelihood inferences of single-copy orthologous sequences (SCOS) (Additional file 1: Fig. S3) and an MSC tree directly from 1.7 million single nucleotide polymorphisms (SNPs), called after mapping available short-read data from nine baleen whale species to the bowhead whale (*Balaena mysticetus*) reference genome (Additional file 1: Fig. S4, [22]. Both approaches resulted in different placements of the gray whale with either low bootstrap support values or higher amounts of quartet

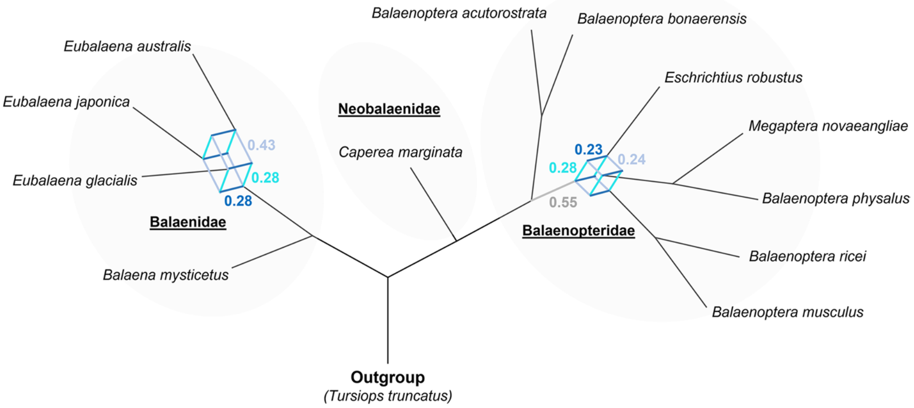

Fig. 2 Consensus network of baleen whale evolution based on whole-genome alignment fragments. The network was constructed from 46,941 fragments of 20 kbp length and a 12% threshold was used to depict conflicts. Extensive phylogenetic conflicts characterize the placement of the gray whale consistent with branch 4 of the main phylogenomic analysis

score conflicts. The consensus phylogeny inferred from SCOS gene trees placed the gray whale at the base of the (fin whale, humpback whale) and (blue whale, rice whale) clades, whereas the MSC phylogeny based on SNPs placed the gray whale together with the blue and rice whale as a sister clade to the fin whale and humpback whale. A consensus network constructed from SCOS data resulted in inconclusive resolution as well (Additional file 1: Fig. S5). Accordingly, we chose the WGA-based topology for all downstream analyses as it depicts the most parsimonious hypothesis with the fewest conflicts in baleen whales.

Divergence time estimates of baleen whales were estimated based on the topology of the consensus WGA tree using branch lengths from an ML analysis of SCOS amino acid data and five calibration points (Fig. 3,

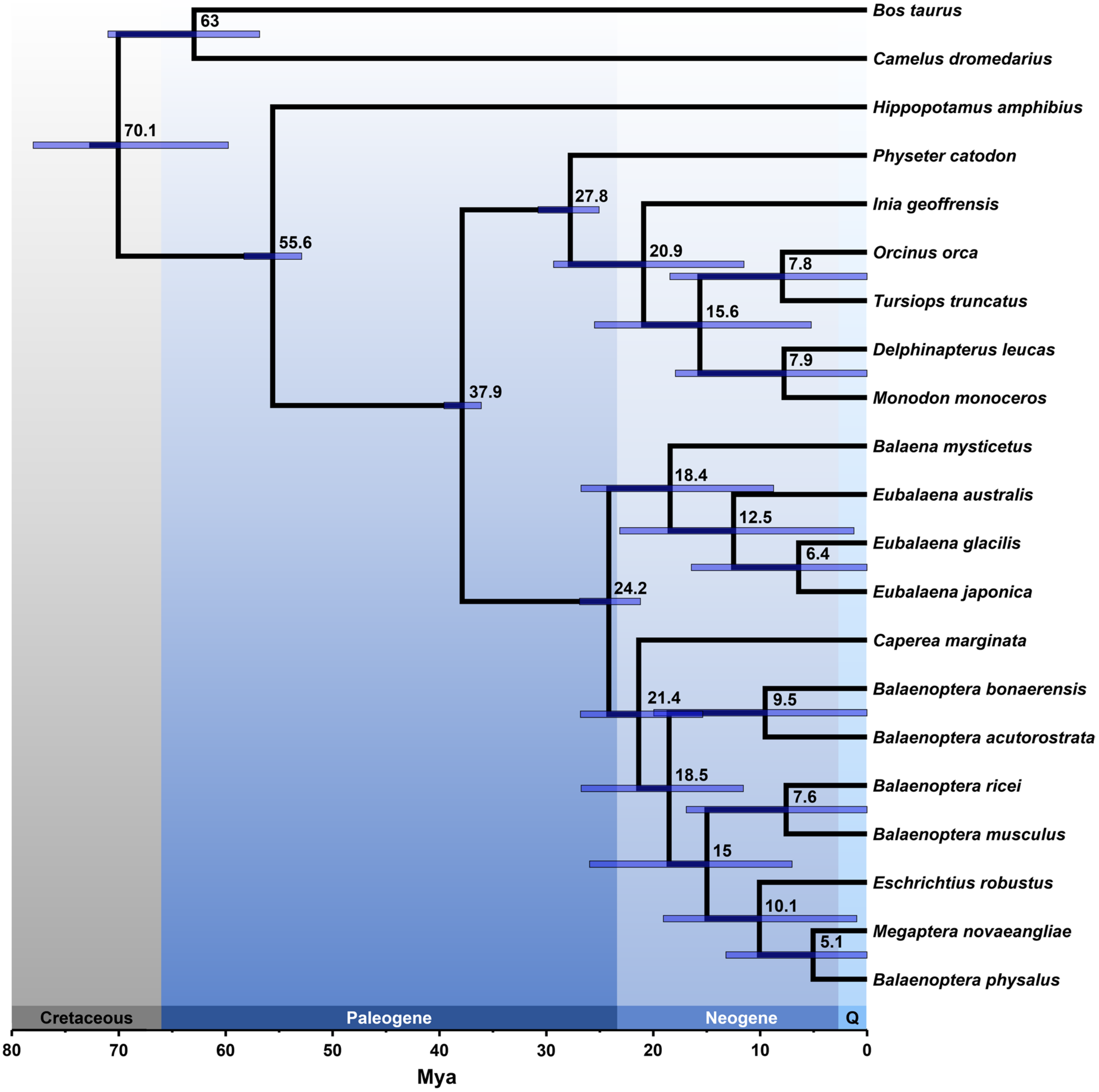

Fig. 3 Divergence time estimates of whales (*Cetacea*). The tree was constructed using the topology presented in Fig. 1, five calibration points (Additional file 1: Table S3 [40–44]), and amino acid sequences of 562 single-copy orthologous sequences. According to this estimate, baleen whales originated at 24.2 Mya, the pygmy right whale diverged around 21.4 Mya, and rorquals around 18.5 Mya, although the error bars indicate large ranges for all three cases

Additional file 1: Table S3 [40–44]). According to these estimates, baleen whales diverged between 26.9 and 21.2 million years ago (Mya), and the split between *Cetotheriidae* and *Balaenopteridae* was estimated to have occurred between 26.8 and 15.4 Mya. Divergence estimated within the rorquals showed a wide range of estimates between 25.9 and 7 Mya.

Demographic inference

The history of the effective population size (Ne) of the pygmy right whale was modeled for a time frame between 10 million years ago (Mya) and 100 thousand years ago (kya) (Fig. 4). The model shows a peak in abundance during the Late-Pleistocene Transition (2.6 Mya) followed by a steady decline until reaching constant numbers after the Mid-Pleistocene Transition (1.6–0.7 Mya) some ~ 400–600 kya. Bootstrap replications closely mirror the initially estimated model, indicating low sampling variance.

Selection analysis of cancer genes

To collect positively selected genes related to body size and hence cancer resistance, we compared rates of non-synonymous (Ka) and synonymous (Ks) substitutions between pairs of large and small-bodied baleen whales. We applied phylogenetic targeting (Additional file 1: Table S4 [45–65], Table S5) and identified three phylogenetically independent pairs that at the same time maximized size differences between them, namely: (1) bowhead whale and pygmy right whale, (2) fin whale and minke whale, and (3) blue whale and rice whale (Fig. 5A). Together with the human reference genome GRCh38, we called single-copy orthologous sequences and collected genes with elevated non-synonymous substitution rates (Ka/Ks > 1) in one or all of the three pairs as a proxy for positive selection. Within our curated set of 1266 single-copy orthologs, we identified 210 unique orthologs with elevated Ka/Ks in at least one of the three pairs. Differentiating between the individual pairs we found: 89 (bowhead/pygmy right), 95 (fin/minke), and 74 (blue/rice) positively selected orthologs, respectively. Six of those genes were found to have an elevated non-synonymous substitution rate in all three pairs (Table 2 [66–70]) that were further functionally specified by BLAST. For five out of six candidate genes, we found evidence in the literature for a correlation between expression patterns and cancer development (Table 2 [66–70]). Furthermore, we found more detailed functional descriptions for two of these genes: first, the C-type lectin domain family 2 member B (CLEC2B/AICL), which is a protein encoded by the natural killer (NK) gene complex proximal CD69 [71]; and second, the RAB15 effector protein (REP15), which together with its associated Rab GTPase controls the flow of transport vesicles in the brain [69]. The ortholog with the highest divergence between Ka and Ks in all three pairs is so far uncharacterized in humans (LOC124907494/LOC124905498) and has only been characterized as "proline-rich" in, e.g., cattle (LOC113892484).

We further characterized enriched functions in the set of 210 positively selected genes by performing a gene enrichment analysis of biological processes against the human reference (Fig. 5B, Table 3). We found 20 enriched Gene Ontology (GO) terms most of which represent general terms related to signaling (GO:0,023,052, GO:0,007,165), cell communication (GO:0,007,154), and

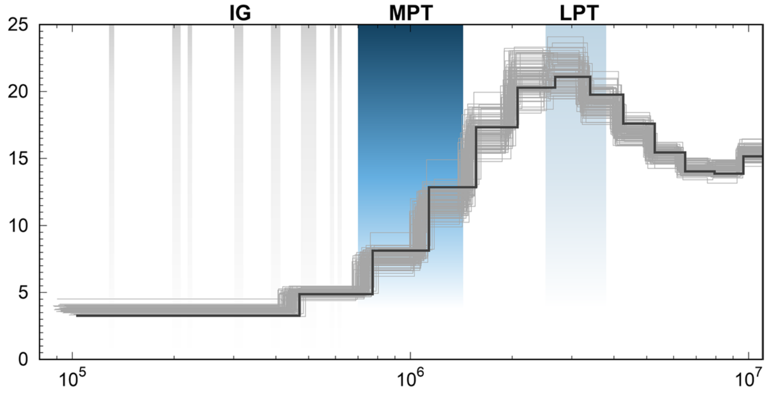

Fig. 4 Demographic history of the pygmy right whale inferred using the PSMC framework. The model covers the last 10 Mya to 100 kya and is scaled based on a mutation rate of 1.38 × 10^{-8} per site per generation [20] and a generation time of 22.1 years, using the minke whale as an approximation for the unknown life expectancy [34]. *x*-axis depicts the time in number of years ago while the *y*-axis depicts the effective population size in thousand individuals. The model indicates a peak of effective population size around the Late-Pleistocene Transition (LPT, 2.6 Mya) (light blue), followed by a steady decline until reaching a lower, but stable population size after the Mid-Pleistocene Transition (MPT, 1 Mya–700 kya) (dark blue). Interglacial periods (gray) did not influence stock sizes

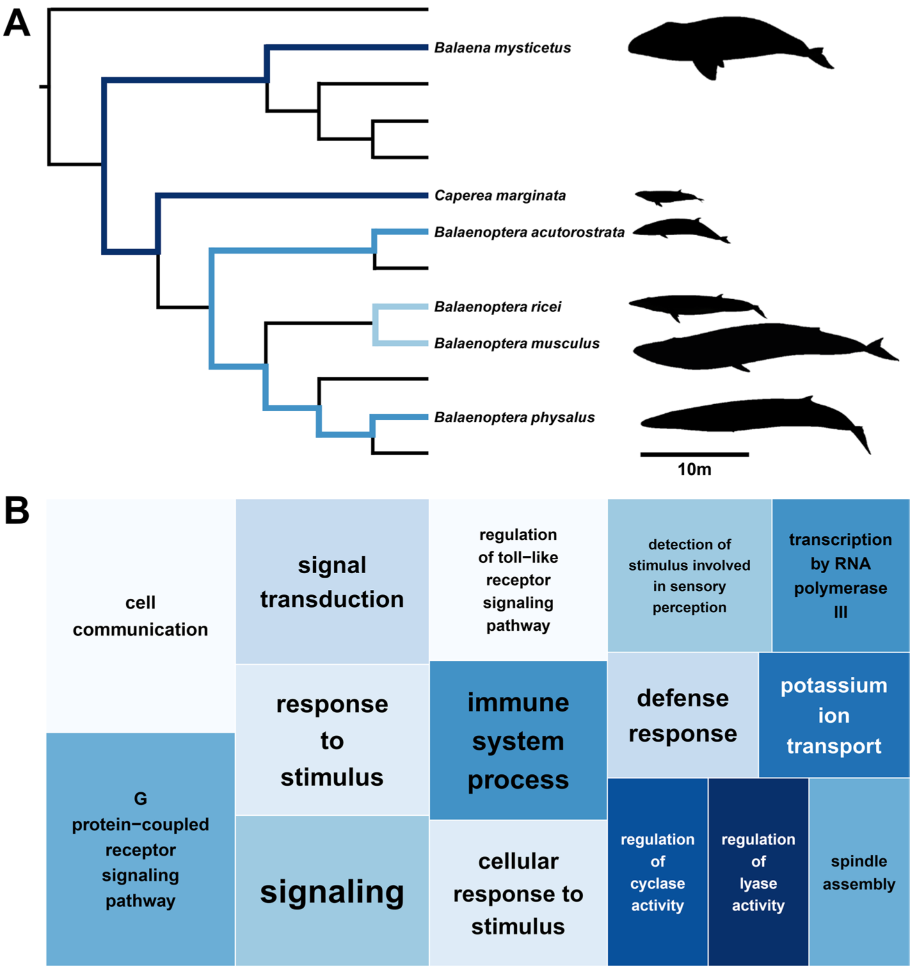
A
Balaena mysticetus
Caperea marginata
Balaenoptera acutorostrata
Balaenoptera ricei
Balaenoptera musculus
Balaenoptera physalus
10m
B
cell communication
signal transduction
regulation of toll−like receptor signaling pathway
detection of stimulus involved in sensory perception
transcription by RNA polymerase III
response to stimulus
immune system process
defense response
potassium ion transport
G protein−coupled receptor signaling pathway
signaling
cellular response to stimulus
regulation of cyclase activity
regulation of lyase activity
spindle assembly

Fig. 5 Enriched functions of genes positively selected in large-bodied baleen whales. **A** Pairs of large and small baleen whales identified using phylogenetic targeting [72] based on the species tree presented in Fig. 1 and length and body mass data (Additional file 1: Table S4 [45–65]). Pairs represent the best combination of phylogenetically independent pairs that maximize size differences, putatively related with cancer resistance (Peto's Paradox, [21]. **B** TreeMap representing gene ontology terms overrepresented in genes with elevated non-synonymous substitution rates in large whales. Rectangle size depicts significance values after false discovery correction after Benjamini and Hochberg

Table 2 List of putative cancer-related candidate genes found to be selected (Ka/Ks > 1) in all three pairwise comparisons of large- and small-bodied baleen whales. Genes were identified based on the imbalance between nonsynonymous (Ka) and synonymous (Ks) substitution rates. Pairs were determined using phylogenetic targeting. Sorted by Ka/Ks ratio. Concerning publications: [66–70]

Symbol	HGNC	Orthogroup	Description	Literature	Ka/Ks	*p*-val
LOC124907494/ LOC124905498	-	OG0018750	Uncharacterized (basic salivary proline-rich protein 1-like in *Bos indicus*)	-	8.19	0.001
C6orf15 (STG)	13,927	OG0018717	Unknown function (expressed in various types of cancer)	Xiong et al., (2022) [66]	2.74	0.165
MGAT4EP	49,418	OG0018311	Pseudogene in human (upregulates the expression of FOXM1 in breast cancer)	Sun et al., (2021) [67]	2.06	0.474
CLEC2B/AICL	2053	OG0018915	Associated with natural killer cells, expressed in various types of cancer	Li et al., (2022) [68]	1.95	0.532
REP15	33,748	OG0018380	Involved in vesicular trafficking, potentially involved in various types of cancer	Rai et al., (2022) [69]	1.67	0.403
C22orf46	26,294	OG0017847	Unknown function (oncogene in adreno-cortical carcinoma)	Li et al., (2020) [70]	1.43	0.435

HGNC, HUGO Gene Nomenclature Committee; *HUGO*, Human Genome Organization

Table 3 List GO terms for biological processes overrepresented in genes positively selected (Ka/Ks > 1) in large-bodied baleen whales. GO terms were collected from the human ortholog of genes with elevated non-synonymous substitution rates in large-bodied whales. GO terms from the human genome (GRCh38) were used as reference. Sorted after significance values

Term ID	Description	Number in reference	Number in query	Fold enrichment	*P*-value*
GO:0,007,186	G-protein-coupled receptor signaling pathway	1098	10	8.1	0.0009
GO:0,007,154	Cell communication	4549	14	2.7	0.0009
GO:0,007,165	Signal transduction	4201	12	2.5	0.0060
GO:0,050,896	Response to stimulus	5812	14	2.2	0.0095
GO:0,023,052	Signaling	4569	12	2.3	0.0095
GO:0,002,376	Immune system process	573	4	6.3	0.0109
GO:0,051,716	Cellular response to stimulus	5003	12	2.1	0.0159
GO:0,006,952	Defense response	257	2	6.9	0.0489

* After false discovery rate correction following Benjamini and Hochberg

response to a stimulus (GO:0,050,896, GO:0,051,716). Terms belonging to the G-protein-coupled receptor signaling pathway (GO:0,007,186) were found to be the most significantly enriched with an 8.9-fold enrichment. We also detected GO terms related to immune system processes (GO:0,002,376) and cell defense (GO:0,006,952). Terms occurring only once were excluded from further discussion.

Discussion

The genome of the pygmy right whale (*Caperea marginata*) allowed us to make a first genetic analysis of a species which is so far nearly unknown to science. Due to its small body size and unique monotypic placement within the mysticetes, the pygmy right whale represents a promising target species to better understand the general evolution of rorquals and to analyze cancer resistance in baleen whales.

Genome features, diversity, and demography

Despite our best efforts to improve the assembly, we did not reach chromosome-level continuity. This could have been caused by either high levels of DNA fragmentation or by a high repeat content, which is typical for baleen whales [73]. Furthermore, there may be unique features in the repetitive sequences of the pygmy right whale genome that further hindered a more continuous assembly. Chromatin-based assembly methods (Hi-C) [74] will likely increase the assembly continuity but rely on fresh tissue samples that are very difficult to get given the elusive nature of the pygmy right whale. Completeness scores also showed that some core genes were fragmented or missing. However, assembly continuity mostly affects the analysis of structural genome changes or gene copy numbers and is therefore unlikely to affect our downstream analyses.

The newly constructed genome allowed us to assess the genetic diversity and to model the previously unknown demographic history. The level of genome-wide heterozygosity is comparable to that of the blue and North Atlantic right whales [75], but higher compared to other mysticetes [20, 75], although comparing genetic diversity alone does not allow conclusions about the well-being of a species [75, 76]. Our demographic model showed a population trajectory over time that is similar to other baleen whales, starting from a high abundance around the Late-Pleistocene Transition (2.6 Mya) and slowly declining in Ne over time until reaching a lower, but stable population size after the Mid-Pleistocene Transition (1.6 Mya–700 kya) [20, 77]. After this point in time, the trajectory shows no indications of an influence from major climatic oscillations that would have consequently affected marine circulation and productivity [78].

Revision of the rorqual phylogeny

The multispecies coalescent (MSC) phylogeny presented in this study is based on fragments of a whole-genome alignment (WGA) that includes data from twelve different whale species including nearly all extant members of the *Mysticeti.* This analysis resulted in an overall well-supported topology that unequivocally placed the pygmy right whale as expected from previous studies [13, 14, 18, 19] but showed a high degree of phylogenetic conflicts for the gray whale (*Eschrichtius robustus*). Other sources of homologous data like, e.g., single-copy orthologous sequences (SCOS) or other methods of gene tree conflation resulted in different placements of the gray whale and even more phylogenetic conflicts across all tree nodes.

These conflicts, depicted by the distribution quartet scores, were found to be nearly even at the branch positioning the gray whale. Because the phylogenomic tree based on WGA fragments had the fewest amount of conflicts in the entire tree, we consider the pairing of the gray whale together with the fin and humpback whale as the most parsimonious explanation of baleen whale evolution, supporting previous findings by [19, 20], and [73]. Nevertheless, we still would not consider this placement as definitive, nor the topology being resolved as a bifurcating event given the nearly equal frequency of conflicting signals. Instead, we suggest that the relationship of the gray whale is best depicted as a polytomy. In the past, such polytomies were thought to be a consequence of a lack of molecular data or taxon sampling and were treated as soft polytomies with the expectation that an increase of data would eventually lead to highly resolved bifurcating trees [79]. However, in this case, only a small increase in the data remains possible and thus we think it is unlikely that the resolution improves in the future. Thus, we consider this polytomy as a hard polytomy that reflects the actual biological history of a rapid radiation of rorquals at the beginning of their divergence 15–25 million years ago (Fig. 3).

Two scenarios can cause sub-trees to deviate from the overall species tree: incomplete lineage sorting (ILS) and introgression. While ILS occurs randomly at predictable frequencies that will not overshadow the true species tree topology [80], introgression might, depending on its extent, obscure the true topology. Within baleen whales, ongoing introgression seems unlikely, given the lack of runs of shared phylogenetic signals across the genome (Fig. 1C). The nearly equal occurring frequencies of alternative topologies in the quartet score analysis (quartet 4, Fig. 1B) would, according to the neutral MSC model, also indicate dominant ILS within the conflicting signals [80]. However, our QuIBL results point out a high amount of ancient introgression (Additional file 1: Table S2) which could have been overshadowed by ILS and recombination over time. These ancient introgression events could also, together with the number of discordant trees caused by ILS, result in a false topology of bifurcating branches that cannot be resolved unless the entire history of introgression events could be unscrambled. Such an attempt to remove putative signals of introgression in a similar case of mbuna cichlids did not alter the distribution of conflicts for the problematic branches [81]. Therefore, choosing between one of the three possible topologies might not be possible, which supports our finding that the evolution of rorquals is best described as a hard polytomy.

Finally, a hard polytomy should not be confused with an unresolved phylogeny, as it is only unresolved in the sense of a strictly bifurcating tree. Yet, with the discover of many hard polytomies like presented here [81–83] we would like to stress that evolution must not be a bifurcating process by all means and that cases like the presented radiation in rorquals are best depicted as an evolutionary network [84].

Identification of cancer-related genes

Our pairwise selection analysis between large and small-bodied mysticetes like the pygmy right whale resulted in a set of 210 candidate genes that might be related to body size and hence cancer resistance following the idea behind Peto's paradox [21]. Within this set, we found six genes with similar signals of positive selection across all three pairs of whales. All except of one were already known to cancer research due to correlations between their over- or under-expression and cancer development (Table 2 [66–70]). The proline-rich protein LOC124907494 (human) is to our knowledge unknown to cancer research and shows the most non-synonymous mutations in large baleen whales, therefore, representing an interesting target for further research. Within our six selected genes, two were already described in greater detail, namely CLEC2B and REP15 [68, 69]. CLEC2B is a member of the C-type lectin domain family 2 and was formally described as activation-induced C-type lectin (AICL) [71]. It is encoded by the natural killer (NK) gene complex proximal CD69 and its transcription is increased during lymphocyte activation [71]. Recently, many studies have highlighted its association with various types of cancer [68, 85] and it is assumed to be connected to the immune response to cancer through ferroptosis activation [68]. Its positive selection in large-bodied baleen whales might represent adaptations to this complex to better control the activation and migration of lymphocytes when encountering cancer. REP15 is the effector of the Rab15 GTPases which are assumed to control vesicular traffic in neuronal tissue [69], a function possibly involved in the adaptions to increase body size. However, a multi-omics analysis also suggested REP15 as a colorectal cancer-specific driving gene [86], and it was shown to interact with other Rab proteins [69] of which many are discussed to be involved in tumorigenesis because of its potential role in signal transduction to stimulate progression and invasion into other areas [87].

Similar implications of adaptions in cell signaling were also found in a gene-enrichment analysis performed on the complete set of 210 candidate genes. Apart of enriched functions involved in signal transduction, we also detected other general GO terms to be enriched like, e.g., cell communication, immune system processes, and defense responses (Table 3). The most enriched function in our analysis was the G-protein-coupled receptor (GPCR) signaling pathway. GPCRs are the largest class of surface-bound receptors with around 900 representatives involved into a variety of basic physiological functions and a growing body of literature describes their diverse implications in cancer initiation, development, survival, and migration [88]. Hence, there is a high potential that adaptions in this pathway resulted in an increased cancer resistance in whales.

Our comparisons between pairs of mysticetes with diverging body sizes revealed that only few genes featured similar signals of positive selection. Furthermore, the here presented enriched functions represent rather general terms that do not allow further specification of exact pathways. One explanation for this might be that large body sizes emerged several times during the evolution of baleen whales, with similar selective pressures towards size increase, but different specific adaptions fixed within the genotypes. Paleontologists have often reported a discrepancy between the size of baleen whale fossil records and extant species [89]. Until 10–12 Mya, mysticetes are considered to have remained less than 10 m long [90], being more comparable to the here featured pygmy right whale than to other gigantic representatives. By combining size records with a phylogenetic framework, Slater et al. (2017) [90] simulated the evolution of large body sizes in baleen whales and located the emergence of gigantism within 5–3 Mya, a period defined by the onset of Northern Hemisphere glaciation and the loss of diversity in small-bodied mysticetes [19]. Increasing positive selection towards larger body sizes may have therefore affected all lineages of baleen whales, resulting in a convergently evolved gigantism within the right whales and multiple clades of rorquals. This evolutionary history would explain the lack of shared candidate genes highlighted in pairwise selection comparisons thus far, including the here presented analysis (Table 2) [22, 24, 25]. Nevertheless, while adaptions in cancer resistance may not be specifically conserved in large-bodied baleen whales, it is still noteworthy that adaptions happened in the same functional categories, because genes belonging to general GO terms like signal transduction, cell communication, immune system processes and cell defense mechanisms were highlighted in every related study thus far [22, 24, 25]. Therefore, focusing efforts to these specific functions might help understanding the whale specific cancer resistance in the future.

Conclusions

In this study, we presented the first de novo genome of the pygmy right whale, the smallest baleen whale species, and the only member of an otherwise extinct family of whales. The genomic data from this species was used to update the baleen whale phylogeny, revealing a hard polytomy between the gray whale and other related rorquals caused by high amounts of introgression at the beginning of their radiation. Additionally, the new genome was included in a genome-wide comparison of selection rates to identify genes related to large body size and hence cancer resistance in mysticetes, resulting in only a small set of common candidate genes supporting a more convergent evolution of gigantism in baleen whales.

Methods

Sampling, DNA isolation, and sequencing

Tissue samples were collected by Prof. Dr. Eric Harley from an individual that was washed ashore at the coast of Simonstown, South Africa, in 1993. Samples were subsequently stored in 70% ethanol at − 20 °C. DNA was extracted from approximately 100 mg of tissue using a standard phenol–chloroform-isoamylalcohol protocol [91]. A 10X Genomics Chromium library was constructed by SciLifeLab and a subsequent sequencing yielded approximately 368.6 million paired/linked 150 bp long Illumina reads (~ 23-fold coverage). To generate long reads, four SMRTbell libraries were constructed following the instructions of the SMRTbell Express Prep kit v2.0 (Pacific Biosciences, Menlo Park, CA). Four SMRT cell sequencing runs were performed in "Continuous Long-Read" (CLR) mode on the Sequel System II with the Sequel II Sequencing Kit 2.0 resulting in ~ 1.9 million reads of 20 kbp or more (~ 35-fold coverage).

Genome assembly and annotation

A whole-genome assembly was performed by Supernova v2.1.1 [92] using the linked short-read data. The intermediate assembly was scaffolded with Sspace-LongRead v1-1 [93] using the long-read data. Gap-closing was performed with TGS-GapCloser v.1.2.0 [94] utilizing the long-read data as well. Polishing was done by first mapping the linked short reads onto the gap-closed assembly with Bowtie 2 v.2.4.5 [95]. The mapping file was filtered for duplicates with the Picard v2.21.2–0 toolkit (https://broadinstitute.github.io/picard/) before being used in variant calling using DeepVariant v.1.3.0 [96]. Resulting variants were utilized to generate a polished assembly by calling consensus sequences with Bcftools v.1.12 [97]. Eventually, gene set completeness was assessed with Busco v5.3.2 [98] by testing the OrthoDB gene sets of *Cetartiodactyla*, *Laurasiatheria*, and *Mammalia* [99].

To test whether the long-read data would result in more continuously assembled contigs when used for the initial assembly, we tested them with different specialized software, namely: Canu v2 [100], wtdbg2 [101], and Flye v2.3.3 [102] following respective user recommendations. We conducted the same downstream efforts to scaffold and polish the resulting assemblies as described above for the final assembly. However, since none of the long-read assemblies reached similar continuity compared to the linked short-read assembly, we eventually decided to use the latter in all following analyses.

Repeats were identified using RepeatModeler v2 (www.repeatmasker.org) and found repeats were merged with the *Cetartiodactyla* repeat database from RepBase [103]. Resulting dataset was used by RepeatMasker v4.1 (www.repeatmasker.org) to mask repeats within the de novo assembly. A first annotation of the genome was performed with the GeMoMa pipeline [104] which identifies genes based on homologous information provided by different annotations of related individuals. Doing so, we collected annotations for other *Cetacea* on various different databases. A complete list of all used annotations can be found in Additional file 1: Table S6 [22, 26, 28–31, 75, 105]. Eventually, found annotations were functionally annotated using InterProScan v5 [32].

Genome diversity and demography

Genome-wide diversity was estimated by the means of genome-wide heterozygosity (HE). Short reads produced by the 10X Chromium platform were therefore trimmed for adapter sequences using the Longranger v.2.2.2 toolkit (https://support.10xgenomics.com) before they were mapped onto the masked de novo assembly with BWA-mem v0.7.17-r1188 (http://bio-bwa.sourceforge.net). Variances were called by BCFtools v1.12 mpileup [97] with the respective "-c" flag and minimal mapping- and base-quality cutoffs of 30. These genotypes were additionally filtered for a too divergent read coverage (> threefold and < 0.3-fold of the expected mean coverage) and for sites with a too high proportion of missing data (5%) using BCFtools v1.12 filter [97]. Eventually, genome-wide heterozygosity was inferred as the proportion of heterozygous genotypes compared to the total genotype set including monomorphic sites.

A first model of the demographic past of the species was constructed with the pairwise sequentially Markovian coalescent (PSMC) framework [33] using the repeat masked genome sequences generated above, a standard of 64 atomic intervals ($-p=4+25*2+4+6$) and a mutation rate of 1.39×10^{-8} per site per generation [24]. Because no generation time estimates exist for the pygmy right whale, we used the generation time of the minke whale as an approximation (22.1 years) [34]. To assess potential variances, 100 bootstrap iterations were performed and were depicted as thinner lines.

Phylogenomics

Phylogenomic reconstruction was conducted based on a whole-genome alignment approach. Therefore, we collected eleven assemblies of other baleen whale species from either the NCBI genome (https://www.ncbi.nlm.nih.gov/genome/) or from DNA Zoo (https://www.dnazoo.org/assemblies). A graphical depiction of our entire phylogenetic workflow is presented in the Additional file 1: Fig. S6. A complete list of all utilized data can be found in Additional file 1: Table S6 [20, 22, 24, 28–31, 75, 105–107].

To generate whole-genome alignments, we followed the overall workflow presented in [108] and most of the respective tools are available on github.com (hillerlab/GenomeAlignmentTools). Briefly, pairwise alignments between the repeat-masked reference genome of the common bottlenose dolphin (*Tursiops truncatus*) (NCBI genome: GCA_011762595.1) and individual baleen whale genomes were constructed using lastz v1.04.15 [35] with the default scoring matrix and the following parameter: "*K* = *2400, L* = *3000, Y* = *9400, H* = *2000*". Co-linear alignment chains were constructed with axtChain [36]. RepeatFiller[37] was used to further align repetitive regions and ChainCleaner [38] with parameters "*LRfoldThreshold* = *2.5 -doPairs -LRfoldThresholdPairs* = *10 -maxPairDistance* = *10,000 -maxSuspectScore* = *100,000 -minBrokenChainScore* = *75,000*" was used to improve alignment specificity. Alignment chains were converted to alignment nets with ChainNet and nets were filtered with NetFilterNonNested.perl, where we applied an overall score threshold of 100,000 and kept syntenic or inverted nets with scores ≥ 5000. Filtered alignment nets were used to compute a whole-genome alignment (WGA) with Multiz-Tba [109] and all unaligned regions were removed from the final alignment.

To generate WGA fragments, we first extracted single species fasta files from the alignment and removed all gaps and ambiguous sites with Bedtools v.2.30 [110]. Bedtools was used to create a dictionary of positions where a gap or ambiguous site occurred in one of the individuals to remove all respective positions in all individuals subsequently. We then generated 20-kbp-sized WGA fragments using the scripts presented in [111]. To evaluate ideal fragment sizes, we conducted an approximately unbiased test (Additional file 1: Fig. S1) [112] as described in Árnason et al. (2018) [20] by testing different placements of the pygmy right whale within the topology of in Árnason et al. (2018) [20]. We further filtered for too conserved and too variable fragments by removing the 5% most variable and least variable fragments, given the maximum likelihood distance inferred using IQTree v.2.1.2 [113]. For each fragment, a phylogenetic tree was constructed using IQTree with 1000 bootstrap replications before summarizing them to a consensus species tree with Astral-III v.5.7.3 [114]. In doing so, we annotated branches with quartet scores and posterior probabilities.

Branch lengths were calculated from a maximum likelihood analysis based on single copy orthologous sequences (SCOS) and a respective pipeline regarding the generation of SCOS datasets and downstream phylogenetic analyses can be found on github.com (magwolf/GEMOMA-TO-PHYLOGENY). Because the resulting set of SCOS is used in multiple downstream analyses (tree calibration and SCOS consensus tree) we included multiple species of *Odontoceti* as well as the hippopotamus (GCA_023065835.1), camel (GCA_000803125.3), and cow (GCA_002263795.3) as outgroups to fulfill the requirements of all analyses. We collected publicly available genome assemblies and protein data as listed in Additional file 1: Table S6. We re-annotated all collected assemblies using the GEMOMA pipeline [104] by conducting homology-based annotations using all available proteomes. SCOS were called by ORTHOFINDER v.2.5.2 [115] using default parameters and the "MSA" method for gene tree inferences. Gene alignments were constructed using MAFFT v.7.475 [116]. To avoid using misaligned or uninformative alignments we applied cutoffs of not more than 40% variable sites and more than 5% variable sites. Alignments were concatenated to a single matrix using FASCONCAT-G v1.04 [117]. The concatenated matrix was trimmed with CLIPKIT v.1.1.3 [118] for informative and conserved sites, allowing an additional gap trimming with the "-m kpic-smart-gap" flag. Eventually, the resulting matrix was used to calculate branch lengths with IQTREE.

Because quartet scores regarding the placement of the gray whale were exceptionally even, we evaluated if they were caused by reginal conflicting sites or if conflicts were evenly distributed across the genome. To do so, all WGA fragments that originated from the reference chromosome_1 were further divided into 1 kbp windows. We assessed their number of informative sites and excluded them when containing less than 50 informative sites. Quartet scores per 20 kbp WGA fragment were then calculated based on the 1kbp windows using IQTREE and ASTRAL-III as described above.

To decipher, whether discordant WGA fragment trees originate from incomplete lineage sorting (ILS) or introgression, we used QUIBL ("quantified introgression via branch lengths") as described in Edelman et al. (2019) [39]. We used 1000 randomly selected trees from the set of filtered 20kbp WGA fragment trees and applied a likelihood threshold of 0.01, 50 EM steps, and a shrinking factor of 0.5. The resulting output can be found in Additional file 1: Table S2. Doing this analysis, QUIBL applies a Bayesian information criterion test (BIC) for both possible scenarios of "ILS" and "ILS + introgression". To decide if a signal truly originates from introgression, we used a strict cutoff of $\Delta BIC < -10$.

Consensus networks were generated using SPLITSTREE 4 [119] and all filtered 20-kbp WGA fragment trees by evaluating different cutoffs of conflicting edges. In doing so, we tested cutoffs between 7 and 30% and eventually used 12% for the final depiction of conflicts.

We further tested the performance of alternative sources of homologous information and alternative construction methods. First, we constructed a consensus tree based on all SCOS trees individually using the workflow described above and evaluated their quartet score distribution across the tree using IQTREE AND ASTRAL-III. Second, we conducted a multispecies coalescent (MSC) tree inference based on single nucleotide polymorphisms (SNPs). These SNPs were called from available short-read data of nine baleen whale species (Additional file 1: Table S6) mapped onto the bowhead whale (*Balaena mysticetus*) reference genome [22] using the same workflow to generate vcf files as described for the pygmy right whale alone during the genetic diversity assessment. The filtered vcf file that still contained monomorphic sites was used to generate biallelic SNPs with VCFTOOLS v.0.1.16 [120]. SNPs were pruned for linkage disequilibrium using the BCFTOOLS plugin "+*prune*" applying a $r2 = 0.9$ cutoff. Coalescence inference was done with SVDQUARTETS [121] which is implemented in PAUP 4.0a (Windows build 169). In doing so, we used the QFM algorithm [122] and conducted 1000 bootstrap replications. A sliding window approach was used to generate a subset of trees with IQTREE using 50 SNPs per window. The set of trees was eventually used for quartet score evaluation of the MSC tree with ASTRAL-III.

A dated tree was constructed by calibrating the topology of the consensus WGA tree presented in the main figure. To include more calibration points, we extended the topology with species of *Odontoceti* as well as hippopotamus, camel, and cow by including the orthologs of all respective species in a SCOS consensus tree as described above. Resulting topology was calibrated with MCMCTREE which is part of the PAML 4.9 package [123] using five calibration points (Additional file 1: Table S3) and the concatenated SCOS matrix.

Selection analyses

Genes putatively involved in cancer resistance in baleen whales were identified by conducting pairwise comparisons of non-synonymous and synonymous substitution rates (Ka/Ks) between a large and a small-bodied baleen whale. To find phylogenetic independent pairs that at the same time maximize size differences within pairs, phylogenetic targeting was applied [72] by specifying the bottlenose dolphin as outgroup and using size and body mass data listed in Additional file 1: Table S4 [45–65]. Candidate pairs were identified based on the tree topology depicted in Fig. 1 and on the standardized summed score.

To collect as many informative orthologs between the six whales as possible, we constructed a second set of SCOS including only the candidate whales as well as the human reference genome GRCh38. We re-run the GEMOMA-TO-PHYLOGENY pipeline, as described above in the phylogenetic section. In doing so, we inferred SCOS between all candidate whales together with the human genome GRCh38. To ensure that alignments were constructed without frameshifts, we first translated nucleotide sequences to amino acid sequences using the EMOSS v6.6.0.0 TRANSEQ [124] tool before generating multiple sequence alignments with MAFFT. Amino acid alignments were then converted back to codon alignments using PAL2NAL v14 [125] using the "-nogap" function to remove gaps as well as in-frame stop codons. To avoid alignment errors being accounted for in downstream Ka/Ks analyses, we removed alignments with the five percent topmost genetic distances using the maximum likelihood distance calculated by IQTREE. Filtered codon alignments were then converted into axt files using AXTCONVERTER [126] before inferring pairwise non-synonymous and synonymous substitution rates with KAKS_CALCULATOR v2 [126]. We screened our results for signals of putative positive selection (Ka/Ks > 1, for the distribution of Ka/Ks see Additional file 1: Fig. S7) and inferred a total as well as a shared set of candidate genes which we further functionally annotated via BLASTn against the general "db" database using a cutoff of $1e^{-25}$.

A gene enrichment analysis was conducted by extracting the human orthologs of all putatively positive selected candidate genes and comparing their functions against the annotation of the human genome reference GRCh38 provided by NCBI (https://www.ncbi.nlm.nih.gov/ projects/genome/guide/human/). Enrichment of functions was inferred using the "parentchild" algorithm implemented in the R-package TOPGO [127]. Significance values were corrected for false discovery rates after Benjamini and Hochberg using the R package p.adjust. We removed GO terms with only a single occurrence to avoid taking artifacts into account. GO terms exceeding a corrected p-value of > 0.05 were eventually used to construct a "TreeMap" with REVIGO [128].

Abbreviations

BIC	Bayesian information criterion
BiK-F	Biodiversity and Climate Research Centre
bp	Base-pairs
BUSCO	Benchmarking Universal Single-Copy Orthologs
CLEC2B	C-type lectin domain family 2 member B
Gbp	Giga-base-pairs
GO	Gene Ontology
GPCR	G-protein-coupled receptor
HE	Heterozygosity
HGNC	HUGO Gene Nomenclature Committee
HUGO	Human Genome Organization
ILS	Incomplete Lineage Sorting
IUCN	International Union for Conservation of Nature
kbp	Kilo-base-pairs
kya	Thousand years ago
Mbp	Mega-base-pairs
ML	Maximum likelihood
MSC	Multi-species coalescent
Mya	Million years ago
NCBI	National Center for Biotechnology Information
PSMC	Pairwise sequentially Markovian coalescent
QuIBL	Quantified introgression events via branch lengths
REP15	RAB15 effector protein
SCOS	Single copy ortholog sequences
SNP	Single nucleotide polymorphism
TBG	Translational Biodiversity Genomics
TSG	Tumor suppressor gene
WGA	Whole-genome alignment

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12915-023-01579-1.

Additional file 1: Fig. S1. An approximately unbiased (AU) test for increasing whole-genome alignment fragment sizes. **Fig. S2.** Consensus networks of baleen whales based on whole-genome alignment fragments and different thresholds. **Fig. S3.** Phylogenomic analysis of baleen whales using protein coding sequences of shared single copy orthologous sequences. **Fig. S4.** Phylogenomic analysis of baleen whales using single nucleotide polymorphisms. **Fig. S5.** Consensus network of Cetacea evolution based on single copy orthologous sequences. **Fig. S6.** Pipeline depicting the process of generating all phylogenomic trees. **Fig. S7.** Distribution of Ka/Ks values over all tested orthologs. **Table S1.** Repeat content of the pygmy right whale assembly. **Table S2.** QuIBL results for all triplets resulting from combining rorquals species affecting the placement of the gray whale. **Table S3.** Calibration points used in the date phylogeny. **Table S4.** Body mass data used for phylogenetic targeting. **Table S5.** Maximal pairs inferred from the phylogenetic targeting analysis. **Table S6**. Used data featured in this study including assemblies, short read archives and proteomes from other *Cetacea or Cetartiodactyla*.

Acknowledgements
We are grateful to Prof. Dr. Eric Harley for his sampling efforts and for providing tissue samples and to Dr. Carola Greve and the TBG Lab Center Team for their assistance with the sequencing process. We thank the Genome Technology Center (RGTC) at Radboudumc for the use of the Sequencing Core Facility (Nijmegen, The Netherlands), which provided the PacBio SMRT sequencing service on the Sequel II platform. Furthermore, we thank Dr. Sven Winter for his assistance in general lab work and Dr. Menno de Jong for bioinformatic support during the selection analysis. We thank Frédérique Lucas for providing the illustration of the pygmy right whale.

Authors' contributions

M.W., U.A., and A.J. conceived and designed the study. All authors wrote the manuscript. M.W. conducted the analyses. K.Z. and D.K.G. helped to construct the final assembly. M.H. constructed the whole-genome alignment. All authors read and approved the final manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The present study was funded by the Centre for Translational Biodiversity Genomics (LOEWE-TBG) through the program "LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education.

Availability of data and materials

Raw sequencing reads have been deposited at the National Center for Biotechnology Information under the BioProject PRJNA856827. The assembled genome sequence of the pygmy right whale is deposited as Genome: JANTQK000000000, BioSample: SAMN29592900. All secondary data, including sequences, alignments, and tree files are uploaded to a Dryad repository:

https://doi.org/10.5061/dryad.9zw3r22j0. The pipeline used to create SCOS datasets and phylogenetic trees is available at: https://doi.org/10.5281/zenodo.7740016 and https://github.com/mag-wolf/GEMOMA-to-Phylogeny. All data generated or analyzed during this study are included in this published article, its supplementary information file, and publicly available repositories. Additional data related to this paper may be requested from the authors.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Senckenberg Biodiversity and Climate Research Centre (BiK-F), Georg-Voigt-Strasse 14-16, Frankfurt Am Main, Germany. [2]Institute for Ecology, Evolution and Diversity, Goethe University, Max-Von-Laue-Strasse. 9, Frankfurt Am Main, Germany. [3]LOEWE-Centre for Translational Biodiversity Genomics (TBG), Senckenberg Nature Research Society, Georg-Voigt-Straße 14-16, Frankfurt Am Main, Germany. [4]Institute of Cell Biology and Neuroscience, Goethe University Frankfurt, Max-Von-Laue-Str. 9, Frankfurt Am Main, Germany. [5]Department of Clinical Sciences Lund, Lund University, Lund, Sweden. [6]Department of Neurosurgery, Skane University Hospital in Lund, Lund, Sweden.

Received: 12 July 2022 Accepted: 27 March 2023
Published online: 12 April 2023

References

1. Kemper CM. Pygmy Right Whale: Caperea marginata. In: Perrin WF, Würsig BG, Thewissen JGM, editors. Encyclopedia of marine mammals. 2nd ed. Amsterdam, Boston, Mass.: Elsevier/Academic Press; 2009. https://doi.org/10.1016/B978-0-12-373553-9.00214-5.
2. Tsai C-H, Mead JG. Crossing the equator: A northern occurrence of the pygmy right whale. Zoological Lett. 2018;4:30. https://doi.org/10.1186/s40851-018-0117-8.
3. Fordyce RE, Marx FG. The pygmy right whale Caperea marginata: The last of the cetotheres. Proc Biol Sci. 2013;280:20122645. https://doi.org/10.1098/rspb.2012.2645.
4. Loch C, Vaz Viegas S, Waddell JN, Kemper C, Cook RB, Werth AJ. Structure and properties of baleen in the Southern right (Eubalaena australis) and Pygmy right whales (Caperea marginata). J Mech Behav Biomed Mater. 2020;110:103939. https://doi.org/10.1016/j.jmbbm.2020.103939.
5. Reeb D, Best PB. Anatomy of the laryngeal apparatus of the Pygmy Right Whale, Caperea marginata (Gray 1846). J Morphol. 1999;242:67–81. https://doi.org/10.1002/(SICI)1097-4687(199910)242:1%3c67::AID-JMOR5%3e3.0.CO;2-%23.
6. Cooke J. Caperea marginata: IUCN Red List of Threatened Species. 2018.
7. Hale HM. The pygmy right whale (Neobalaena marginata) in South Australian waters, Part 2. Records South Aust Mus. 1964;14:679–94.
8. Rice DW. Marine mammals of the world: Systematics and distribution / Dale W. Rice. Lawrence, KS: Society for Marine Mammalogy. 1998.
9. Steeman ME. Cladistic analysis and a revised classification of fossil and recent mysticetes. Zool J Linn Soc. 2007;150:875–94. https://doi.org/10.1111/j.1096-3642.2007.00313.x.
10. Ekdale EG, Berta A, Deméré TA. The comparative osteology of the petrotympanic complex (ear region) of extant baleen whales (Cetacea: Mysticeti). PLoS One. 2011;6:e21311. https://doi.org/10.1371/journal.pone.0021311.
11. Churchill M, Berta A, Deméré T. The systematics of right whales (Mysticeti: Balaenidae). Mar Mamm Sci. 2012;28:497–521. https://doi.org/10.1111/j.1748-7692.2011.00504.x.
12. Árnason Ú, Best PB. Phylogenetic relationships within the Mysticeti (whalebone whales) based upon studies of highly repetitive DNA in all extant species. Hereditas. 1991;114:263–9. https://doi.org/10.1111/j.1601-5223.1991.tb00333.x.
13. Árnason Ú, Gullberg A. Cytochrome b nucleotide sequences and the identification of five primary lineages of extant cetaceans. Mol Biol Evol. 1996;13:407–17. https://doi.org/10.1093/oxfordjournals.molbev.a025599.
14. Árnason Ú, Gullberg A, Janke A. Mitogenomic analyses provide new insights into cetacean origin and evolution. Gene. 2004;333:27–34. https://doi.org/10.1016/j.gene.2004.02.010.
15. Deméré TA, McGowen MR, Berta A, Gatesy J. Morphological and molecular evidence for a stepwise evolutionary transition from teeth to baleen in mysticete whales. Syst Biol. 2008;57:15–37. https://doi.org/10.1080/10635150701884632.
16. McGowen MR, Spaulding M, Gatesy J. Divergence date estimation and a comprehensive molecular tree of extant cetaceans. Mol Phylogenet Evol. 2009;53:891–906. https://doi.org/10.1016/j.ympev.2009.08.018.
17. Steeman ME, Hebsgaard MB, Fordyce RE, Ho SYW, Rabosky DL, Nielsen R, et al. Radiation of extant cetaceans driven by restructuring of the oceans. Syst Biol. 2009;58:573–85. https://doi.org/10.1093/sysbio/syp060.
18. McGowen MR, Tsagkogeorga G, Álvarez-Carretero S, Dos Reis M, Struebig M, Deaville R, et al. Phylogenomic Resolution of the Cetacean Tree of Life Using Target Sequence Capture. Syst Biol. 2020;69:479–501. https://doi.org/10.1093/sysbio/syz068.
19. Marx FG, Fordyce RE. Baleen boom and bust: A synthesis of mysticete phylogeny, diversity and disparity. R Soc Open Sci. 2015;2:140434. https://doi.org/10.1098/rsos.140434.
20. Árnason Ú, Lammers F, Kumar V, Nilsson MA, Janke A. Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. Sci Adv. 2018;4:eaap9873. https://doi.org/10.1126/sciadv.aap9873.
21. Peto R, Roe FJ, Lee PN, Levy L, Clack J. Cancer and ageing in mice and men. Br J Cancer. 1975;32:411–26. https://doi.org/10.1038/bjc.1975.242.
22. Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, et al. Insights into the evolution of longevity from the bowhead whale genome. Cell Rep. 2015;10:112–22. https://doi.org/10.1016/j.celrep.2014.12.008.
23. Nunney L, Maley CC, Breen M, Hochberg ME, Schiffman JD. Peto's paradox and the promise of comparative oncology. Philos Trans R Soc Lond B Biol Sci. 2015. https://doi.org/10.1098/rstb.2014.0177.
24. Tollis M, Robbins J, Webb AE, Kuderna LFK, Caulin AF, Garcia JD, et al. Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). Mol Biol Evol. 2019;36:1746–63. https://doi.org/10.1093/molbev/msz099.
25. Jossey S, Haddrath O, Loureiro L, Lim B, Miller J, Lok S, et al. Blue whale (Balaenoptera musculus musculus) genome: Population structure and history in the North Atlantic. 2021.
26. Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet. 2014;46:88–92. https://doi.org/10.1038/ng.2835.
27. Venkat A, Hahn MW, Thornton JW. Multinucleotide mutations cause false inferences of lineage-specific positive selection. Nat Ecol Evol. 2018;2:1280–8. https://doi.org/10.1038/s41559-018-0584-5.
28. Elbers JP, Rogers MF, Perelman PL, Proskuryakova AA, Serdyukova NA, Johnson WE, et al. Improving Illumina assemblies with Hi-C and long reads: An example with the North African dromedary. Mol Ecol Resour. 2019;19:1015–26. https://doi.org/10.1111/1755-0998.13020.
29. Westbury MV, Petersen B, Garde E, Heide-Jørgensen MP, Lorenzen ED. Narwhal Genome Reveals Long-Term Low Genetic Diversity despite Current Large Abundance Size. iScience. 2019;15:592–9. https://doi.org/10.1016/j.isci.2019.03.023.
30. Foote AD, Liu Y, Thomas GWC, Vinař T, Alföldi J, Deng J, et al. Convergent evolution of the genomes of marine mammals. Nat Genet. 2015;47:272–5. https://doi.org/10.1038/ng.3198.
31. Fan G, Zhang Y, Liu X, Wang J, Sun Z, Sun S, et al. The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution. Mol Ecol Resour. 2019;19:944–56. https://doi.org/10.1111/1755-0998.13003.
32. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale protein function classification. Bioinformatics. 2014;30:1236–40. https://doi.org/10.1093/bioinformatics/btu031.

33. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6. https://doi.org/10.1038/nature10231.
34. Taylor BL, Chivers S, Larese J, Perrin WF. Generation length and percent mature estimates for IUCN assessments of cetaceans. Southwest Fisheries Sci Center Adm Rep. 2007;10:LJ-07-01.
35. Harris RS. Improved Pairwise Alignment of Genomic DNA. [Ph.D.]. Pennsylvania, USA: Pennsylvania State University; 2007.
36. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. Proc Natl Acad Sci U S A. 2003;100:11484–9. https://doi.org/10.1073/pnas.1932072100.
37. Osipova E, Hecker N, Hiller M. RepeatFiller newly identifies megabases of aligning repetitive sequences and improves annotations of conserved non-exonic elements. Gigascience. 2019. https://doi.org/10.1093/gigascience/giz132.
38. Suarez HG, Langer BE, Ladde P, Hiller M. chainCleaner improves genome alignment specificity and sensitivity. Bioinformatics. 2017;33:1596–603. https://doi.org/10.1093/bioinformatics/btx024.
39. Edelman NB, Frandsen PB, Miyagi M, Clavijo B, Davey J, Dikow RB, et al. Genomic architecture and introgression shape a butterfly radiation. Science. 2019;366:594–9. https://doi.org/10.1126/science.aaw2090.
40. Gingerich PD. New Earliest Wasatchian Mammalian Fauna from the Eocene of Northwestern Wyoming: Composition and Diversity in a Rarely Sampled High-Floodplain Assemblage. 1989.
41. Bajpai S, Gingerich PD. A new Eocene archaeocete (Mammalia, Cetacea) from India and the time of origin of whales. Proc Natl Acad Sci U S A. 1998;95:15464–8. https://doi.org/10.1073/pnas.95.26.15464.
42. Lambert O, Martínez-Cáceres M, Bianucci G, Di Celma C, Salas-Gismondi R, Steurbaut E, et al. Earliest Mysticete from the Late Eocene of Peru Sheds New Light on the Origin of Baleen Whales. Curr Biol. 2017;27:1535-1541.e2. https://doi.org/10.1016/j.cub.2017.04.026.
43. Bisconti M. Skull morphology and phylogenetic relationships of a new diminutive balaenid from the lower pliocene of Belgium. Palaeontology. 2005;48:793–816. https://doi.org/10.1111/j.1475-4983.2005.00488.x.
44. Boersma AT, Pyenson ND. Arktocara yakataga, a new fossil odontocete (Mammalia, Cetacea) from the Oligocene of Alaska and the antiquity of Platanistoidea. PeerJ. 2016;4:e2321. https://doi.org/10.7717/peerj.2321.
45. Budylenko GA, Panfilov BG, Pakhomova AA, Sazhinov EG. New data on pygmy right whales Neobalaena marginata (Gray, 1848). Trudy Atlanticheskii Nauchno-Issledovatel'skii Institut Rybnogo Khozyaistva I Okeanografii. 1973;51:122–32.
46. George JC, Bada J, Zeh J, Scott L, Brown SE, O'Hara T, Suydam R. Age and growth estimates of bowhead whales (Balaena mysticetus) via aspartic acid racemization. Can J Zool. 1999;77:571–80. https://doi.org/10.1139/z99-015.
47. Hamilton PK, Knowlton AR, Marx MK, Kraus SD. Age structure and longevity in North Atlantic right whales Eubalaena glacialis and their relation to reproduction. Mar Ecol Prog Ser. 1998;171:285–92. https://doi.org/10.3354/meps171285.
48. Fortune SME, Moore MJ, Perryman WL, Trites AW. Body growth of North Atlantic right whales (Eubalaena glacialis) revisited. Mar Mamm Sci. 2021;37:433–47. https://doi.org/10.1111/mms.12753.
49. Christiansen F, Sironi M, Moore MJ, Di Martino M, Ricciardi M, Warick HA, et al. Estimating body mass of free-living whales using aerial photogrammetry and 3D volumetrics. Methods Ecol Evol. 2019;10:2034–44. https://doi.org/10.1111/2041-210X.13298.
50. Lockyer C. Body weights of some species of large whales. ICES J Mar Sci. 1976;36:259–73. https://doi.org/10.1093/icesjms/36.3.259.
51. Konishi K. Characteristics of blubber distribution and body condition indicators for Antarctic minke whales (Balaenoptera bonaerensis). Mammal Study. 2006;31:15–22. https://doi.org/10.3106/1348-6160(2006)31[15:COBDAB]2.0.CO;2.
52. Markussen NH, Ryg M, Lydersen C. Food consumption of the NE Atlantic minke whale (Balaenoptera acutorostrata) population estimated with a simulation model. ICES J Mar Sci. 1992;49:317–23. https://doi.org/10.1093/icesjms/49.3.317.
53. Horwood J. Biology and exploitation of the minke whale. Boca Raton, Fla: CRC Press; 1990.
54. Ruud JT. The Blue Whale. Sci Am. 1956;195:46–51.
55. Gilpatrick JW, Perryman WL. Geographic variation in external morphology of North Pacific and Southern Hemisphere blue whales (Balaenoptera musculus). J Cetac Res Manage. 2008;10:9–21.
56. Sears R, Perrin WF. Blue Whale: Balaenoptera musculus. In: Perrin WF, Würsig BG, Thewissen JGM, editors. Encyclopedia of marine mammals. 2nd ed. Amsterdam, Boston, Mass: Elsevier/Academic Press; 2009. p. 120–4.
57. Tershy BR. Body Size, Diet, Habitat Use, and Social Behavior of Balaenoptera Whales in the Gulf of California. J Mammal. 1992;73:477–86. https://doi.org/10.2307/1382013.
58. Rosel PE, Wilcox LA, Yamada TK, Mullin KD. A new species of baleen whale (Balaenoptera) from the Gulf of Mexico, with a review of its geographic distribution. Mar Mamm Sci. 2021;37:577–610. https://doi.org/10.1111/mms.12776.
59. Rice DW, Wolman AA. The life history and ecology of the gray whale (Eschrichtius robustus): American Society of Mammalogist. 1971.
60. Swartz SL. Gray Whale: Eschrichtius robustus. In: Würsig BG, Thewissen JGM, Kovacs KM, editors. Encyclopedia of marine mammals. Amsterdam: Academic Press; 2017. p. 422–8.
61. Chittleborough RG. Determination of age in the humpback whale, Megaptera nodosa (Bonnaterre). Mar Freshw Res. 1959;10:125–43.
62. Jefferson TA, Pitman RL, Webber MA, editors. Marine mammals of the world: A comprehensive guide to their identification. 2nd ed. Amsterdam: Academic Press; 2015.
63. Clapham PJ, Mead JG. Megaptera novaeangliae. Mammalian Species. 1999:1.https://doi.org/10.2307/3504352.
64. Lockyer C, Waters T. Weights and anatomical measurements of northeastern atlantic fin (balaenoptera physalus, linnaeus) and sei (B. Borealis, lesson) whales. Marine Mammal Sci. 1986;2:169–85. https://doi.org/10.1111/j.1748-7692.1986.tb00039.x.
65. Aguilar A, García-Vernet R. Fin whale: Balaenoptera physalus. In: Würsig BG, Thewissen JGM, Kovacs KM, editors. Encyclopedia of marine mammals. Amsterdam: Academic Press; 2017. p. 368–71.
66. Xiong X, Wang S, Ye Y, Gao Z. C6orf15 acts as a potential novel marker of adverse pathological features and prognosis for colon cancer. 2022.
67. Sun M, Wang Y, Zheng C, Wei Y, Hou J, Zhang P, et al. Systematic functional interrogation of human pseudogenes using CRISPRi. Genome Biol. 2021;22:240. https://doi.org/10.1186/s13059-021-02464-2.
68. Li X, Tao X, Ding X. An integrative analysis to reveal that CLEC2B and ferroptosis may bridge the gap between psoriatic arthritis and cancer development. Sci Rep. 2022;12:14653. https://doi.org/10.1038/s41598-022-19135-2.
69. Rai A, Singh AK, Bleimling N, Posern G, Vetter IR, Goody RS. Rep15 interacts with several Rab GTPases and has a distinct fold for a Rab effector. Nat Commun. 2022;13:4262. https://doi.org/10.1038/s41467-022-31831-1.
70. Li W, Liu R, Wei D, Zhang W, Zhang H, Huang W, Hao L. Circular RNA circ-CCAC1 Facilitates Adrenocortical Carcinoma Cell Proliferation, Migration, and Invasion through Regulating the miR-514a-5p/C22orf46 Axis. Biomed Res Int. 2020;2020:3501451. https://doi.org/10.1155/2020/3501451.
71. Hamann J, Montgomery KT, Lau S, Kucherlapati R, van Lier RA. AICL: A new activation-induced antigen encoded by the human NK gene complex. Immunogenetics. 1997;45:295–300. https://doi.org/10.1007/s002510050208.
72. Arnold C, Nunn CL. Phylogenetic targeting of research effort in evolutionary biology. Am Nat. 2010;176:601–12.
73. Lammers F, Blumer M, Rücklé C, Nilsson MA. Retrophylogenomics in rorquals indicate large ancestral population sizes and a rapid radiation. Mob DNA. 2019;10:5. https://doi.org/10.1186/s13100-018-0143-2.
74. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J. Hi-C: A comprehensive technique to capture the conformation of genomes. Methods. 2012;58:268–76. https://doi.org/10.1016/j.ymeth.2012.05.001.
75. Wolf M, de Jong M, Halldórsson SD, Árnason Ú, Janke A. Genomic impact of whaling in North Atlantic Fin Whales. Mol Biol Evol. 2022. https://doi.org/10.1093/molbev/msac094.
76. Teixeira JC, Huber CD. The inflated significance of neutral genetic diversity in conservation genetics. Proc Natl Acad Sci U S A. 2021. https://doi.org/10.1073/pnas.2015096118.
77. Cerca J, Westbury MV, Heide-Jørgensen MP, Kovacs KM, Lorenzen ED, Lydersen C, et al. High genomic diversity in the endangered

East Greenland Svalbard Barents Sea stock of bowhead whales (Balaena mysticetus). Sci Rep. 2022;12:6118. https://doi.org/10.1038/s41598-022-09868-5.
78. Sigman DM, Hain MP, Haug GH. The polar ocean and glacial cycles in atmospheric CO(2) concentration. Nature. 2010;466:47–55. https://doi.org/10.1038/nature09149.
79. Evolution GH. Ending incongruence. Nature. 2003;425:782. https://doi.org/10.1038/425782a.
80. Hibbins MS, Hahn MW. Phylogenomic approaches to detecting and characterizing introgression. Genetics. 2022. https://doi.org/10.1093/genetics/iyab173.
81. Scherz MD, Masonick P, Meyer A, Hulsey CD. Between a Rock and a Hard Polytomy: Phylogenomics of the Rock-Dwelling Mbuna Cichlids of Lake Malaŵi. Syst Biol. 2022;71:741–57. https://doi.org/10.1093/sysbio/syac006.
82. Suh A. The phylogenomic forest of bird trees contains a hard polytomy at the root of Neoaves. Zool Scr. 2016;45:50–62. https://doi.org/10.1111/zsc.12213.
83. Nilsson MA, Zheng Y, Kumar V, Phillips MJ, Janke A. Speciation Generates Mosaic Genomes in Kangaroos. Genome Biol Evol. 2018;10:33–44. https://doi.org/10.1093/gbe/evx245.
84. Bapteste E, van Iersel L, Janke A, Kelchner S, Kelk S, McInerney JO, et al. Networks: Expanding evolutionary thinking. Trends Genet. 2013;29:439–41. https://doi.org/10.1016/j.tig.2013.05.007.
85. Tang S, Huang X, Jiang H, Qin S. Identification of a Five-Gene Prognostic Signature Related to B Cells Infiltration in Pancreatic Adenocarcinoma. Int J Gen Med. 2021;14:5051–68. https://doi.org/10.2147/IJGM.S324432.
86. Xu X, Gong C, Wang Y, Hu Y, Liu H, Fang Z. Multi-omics analysis to identify driving factors in colorectal cancer. Epigenomics. 2020;12:1633–50. https://doi.org/10.2217/epi-2020-0073.
87. Tzeng H-T, Wang Y-C. Rab-mediated vesicle trafficking in cancer. J Biomed Sci. 2016;23:70. https://doi.org/10.1186/s12929-016-0287-7.
88. Chaudhary PK, Kim S. An Insight into GPCR and G-Proteins as Cancer Drivers. Cells. 2021. https://doi.org/10.3390/cells10123288.
89. Tsai C-H, Kohno N. Multiple origins of gigantism in stem baleen whales. Naturwissenschaften. 2016;103:89. https://doi.org/10.1007/s00114-016-1417-5.
90. Slater GJ, Goldbogen JA, Pyenson ND. Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. Proc Biol Sci. 2017. https://doi.org/10.1098/rspb.2017.0546.
91. Sambrook J, Russell DW. Isolation of High-molecular-weight DNA from Mammalian Cells Using Formamide. CSH Protoc. 2006. https://doi.org/10.1101/pdb.prot3225.
92. Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. Direct determination of diploid genome sequences. Genome Res. 2017;27:757–67. https://doi.org/10.1101/gr.214874.116.
93. Boetzer M, Pirovano W. SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 2014;15:211. https://doi.org/10.1186/1471-2105-15-211.
94. Xu M, Guo L, Gu S, Wang O, Zhang R, Peters BA, et al. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. Gigascience. 2020. https://doi.org/10.1093/gigascience/giaa094.
95. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9. https://doi.org/10.1038/nmeth.1923.
96. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36:983–7. https://doi.org/10.1038/nbt.4235.
97. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021. https://doi.org/10.1093/gigascience/giab008.
98. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647–54. https://doi.org/10.1093/molbev/msab199.
99. Zdobnov EM, Kuznetsov D, Tegenfeldt F, Manni M, Berkeley M, Kriventseva EV. OrthoDB in 2020: Evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2021;49:D389–93. https://doi.org/10.1093/nar/gkaa1009.
100. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36. https://doi.org/10.1101/gr.215087.116.
101. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17:155–8. https://doi.org/10.1038/s41592-019-0669-3.
102. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37:540–6. https://doi.org/10.1038/s41587-019-0072-8.
103. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7. https://doi.org/10.1159/000084979.
104. Keilwagen J, Hartung F, Grau J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. Methods Mol Biol. 2019;1962:161–77. https://doi.org/10.1007/978-1-4939-9173-0_9.
105. Jones SJM, Taylor GA, Chan S, Warren RL, Hammond SA, Bilobram S, et al. The Genome of the Beluga Whale (Delphinapterus leucas). Genes (Basel). 2017. https://doi.org/10.3390/genes8120378.
106. Kishida T, Thewissen J, Hayakawa T, Imai H, Agata K. Aquatic adaptation and the evolution of smell and taste in whales. Zoological Lett. 2015;1:9. https://doi.org/10.1186/s40851-014-0002-z.
107. Yuan Y, Zhang Y, Zhang P, Liu C, Wang J, Gao H, et al. Comparative genomics provides insights into the aquatic adaptations of mammals. Proc Natl Acad Sci U S A. 2021. https://doi.org/10.1073/pnas.2106080118.
108. Hecker N, Hiller M. A genome alignment of 120 mammals highlights ultraconserved element variability and placenta-associated enhancers. Gigascience. 2020. https://doi.org/10.1093/gigascience/giz159.
109. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004;14:708–15. https://doi.org/10.1101/gr.1933104.
110. Quinlan AR. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. Curr Protoc Bioinformatics. 2014;47:11.12.1-34. https://doi.org/10.1002/0471250953.bi1112s47.
111. Coimbra RTF, Winter S, Kumar V, Koepfli K-P, Gooley RM, Dobrynin P, et al. Whole-genome analysis of giraffe supports four distinct species. Curr Biol. 2021;31:2929-2938.e5. https://doi.org/10.1016/j.cub.2021.04.033.
112. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002;51:492–508. https://doi.org/10.1080/10635150290069913.
113. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37:1530–4. https://doi.org/10.1093/molbev/msaa015.
114. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19:153. https://doi.org/10.1186/s12859-018-2129-y.
115. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20:238. https://doi.org/10.1186/s13059-019-1832-y.
116. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018;34:2490–2. https://doi.org/10.1093/bioinformatics/bty121.
117. Kück P, Longo GC. FASconCAT-G: Extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Front Zool. 2014;11:81. https://doi.org/10.1186/s12983-014-0081-x.
118. Steenwyk JL, Buida TJ, Li Y, Shen X-X, Rokas A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 2020;18:e3001007. https://doi.org/10.1371/journal.pbio.3001007.
119. Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 2006;23:254–67. https://doi.org/10.1093/molbev/msj030.

120. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8. https://doi.org/10.1093/bioinformatics/btr330.
121. Chifman J, Kubatko L. Quartet inference from SNP data under the coalescent model. Bioinformatics. 2014;30:3317–24. https://doi.org/10.1093/bioinformatics/btu530.
122. Reaz R, Bayzid MS, Rahman MS. Accurate phylogenetic tree reconstruction from quartets: A heuristic approach. PLoS One. 2014;9:e104008. https://doi.org/10.1371/journal.pone.0104008.
123. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91. https://doi.org/10.1093/molbev/msm088.
124. Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. 2022. https://doi.org/10.1093/nar/gkac240.
125. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34:W609–12. https://doi.org/10.1093/nar/gkl315.
126. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. Genom Proteomics Bioinform. 2010;8:77–80. https://doi.org/10.1016/S1672-0229(10)60008-3.
127. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 2006;22:1600–7. https://doi.org/10.1093/bioinformatics/btl140.
128. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6:e21800. https://doi.org/10.1371/journal.pone.0021800.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary Materials for

The genome of the pygmy right whale illuminates the evolution of rorquals.

Magnus Wolf[1,2], Konstantin Zapf[1,2], Deepak Kumar Gupta[3], Michael Hiller[1,3], Úlfur Árnason[4,5], Axel Janke[1,2,3]

*Corresponding author: Magnus Wolf; Email: Magnus.Wolf@senckenberg.de

This PDF file includes:

Figs. S1 to S7

Tables S1 to S6

Supplementary Figures

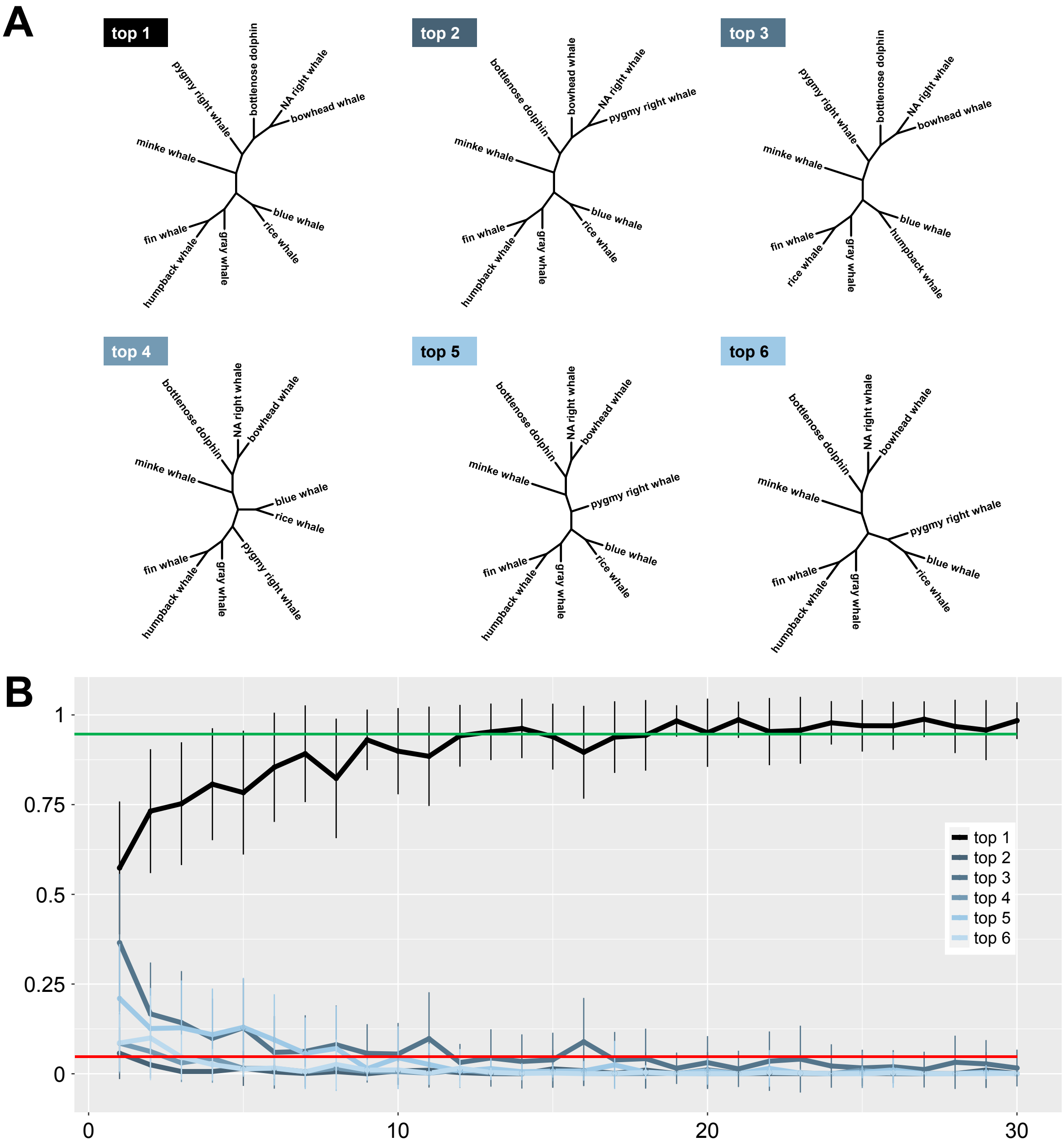

Fig. S1 An approximately unbiased (AU) test for increasing whole-genome alignment fragment sizes. A Different topologies (top1-top6) evaluated by testing different placements of the pygmy right whale (*Caperea marginata*). **B** Distribution of AU values for increasing fragment sizes given different topologies. The green and red line mark pAU=0.05 intervals at which alternative hypothesis could be rejected. Based on this AU test, a fragment size of 20 kbp was chosen for downstream analyses.

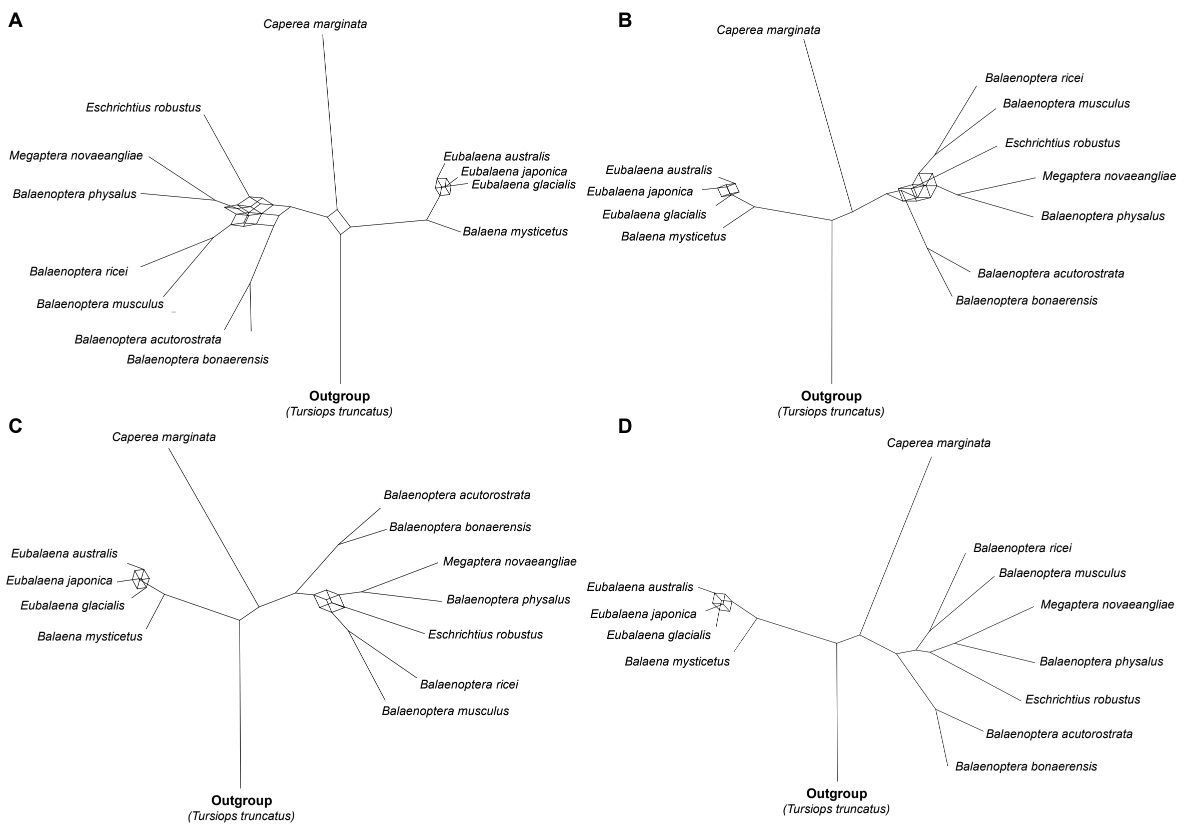

Fig. S2 Consensus networks of baleen whales based on whole-genome alignment fragments and different thresholds. Thresholds were lowered stepwise starting from 30% until reaching 5%. **A**. 5% - 7%, **B**. 7% - 11%, **C**. 11% - 12%, **D**. 12% - 30%.

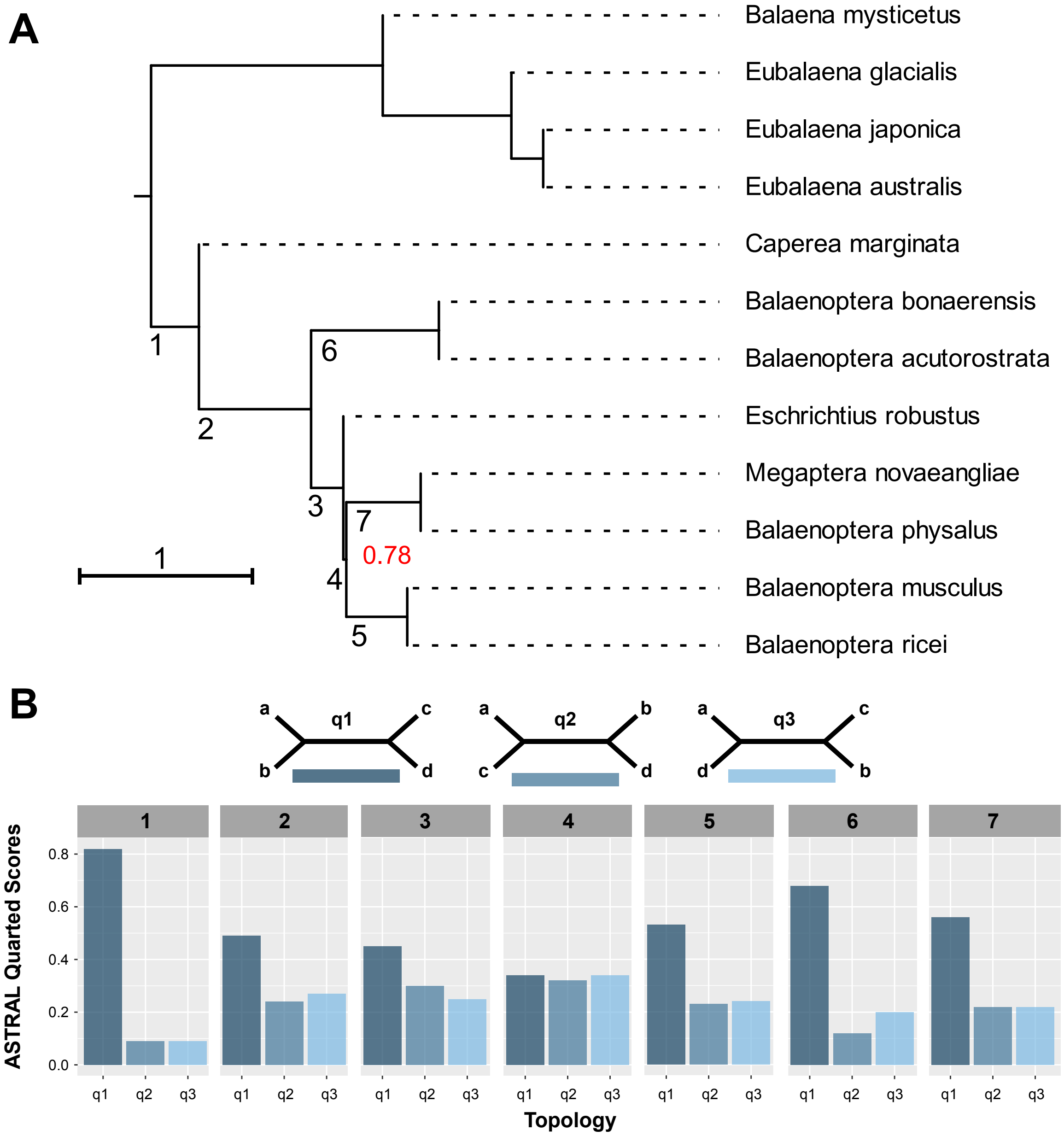

Fig. S3 Phylogenomic analysis of baleen whales using protein coding sequences of shared single copy orthologous sequences. **A**. Phylogenomic multi-species coalescent (MSC) tree conflated from 563 trees that were each constructed from SCOS. All branches except the branch separating (fin whale and humpback whale) and (blue whale and rice whale) received maximum bootstrap support while the latter received 78% support. The pygmy right whale was placed at the base of the rorquals, and the gray whale was placed with a short branch at the base of other large rorquals. **B**. Quartet scores of different branches across the MSC tree. Branches 1 - 7 were analyzed for the amount of gene trees supporting one of the three possible unrooted topologies (q1 - q3). All branches featured substantial conflicting signals while branch 4 received nearly equal frequencies for each alternative topology similar to what was shown in Fig 1 of the main manuscript.

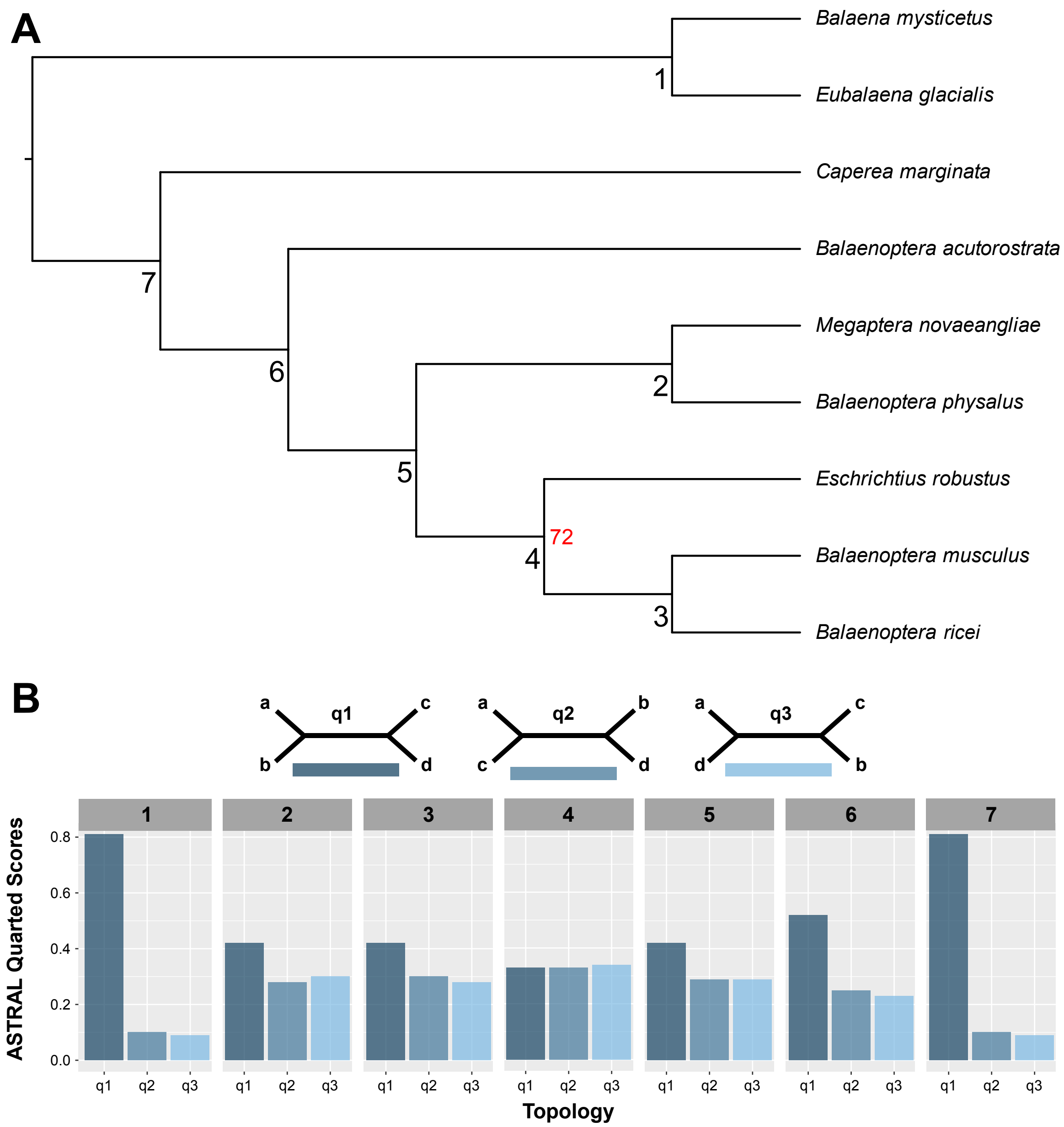

Fig. S4 Phylogenomic analysis of baleen whales using single nucleotide polymorphisms. **A**. Phylogenomic multi-species coalescent (MSC) tree directly inferred from 1.7 million SNPs. All branches except the branch grouping the gray whale with the blue and rice whale received 100% bootstrap support while the latter received 72%. The pygmy right whale was placed at the base of the rorquals, and the gray whale was grouped together with the blue and rice whale forming sister clade to the fin and humpback whale. **B**. Quartet scores of different branches across the MSC tree were inferred using trees constructed from 50 SNP windows. Branches 1 - 7 were analyzed for the amount of trees supporting one of the three possible unrooted topologies (q1 - q3). All branches featured substantial conflicting signals while branch 4 received nearly equal frequencies for each alternative topology similar to what was shown in Fig 1 of the main manuscript.

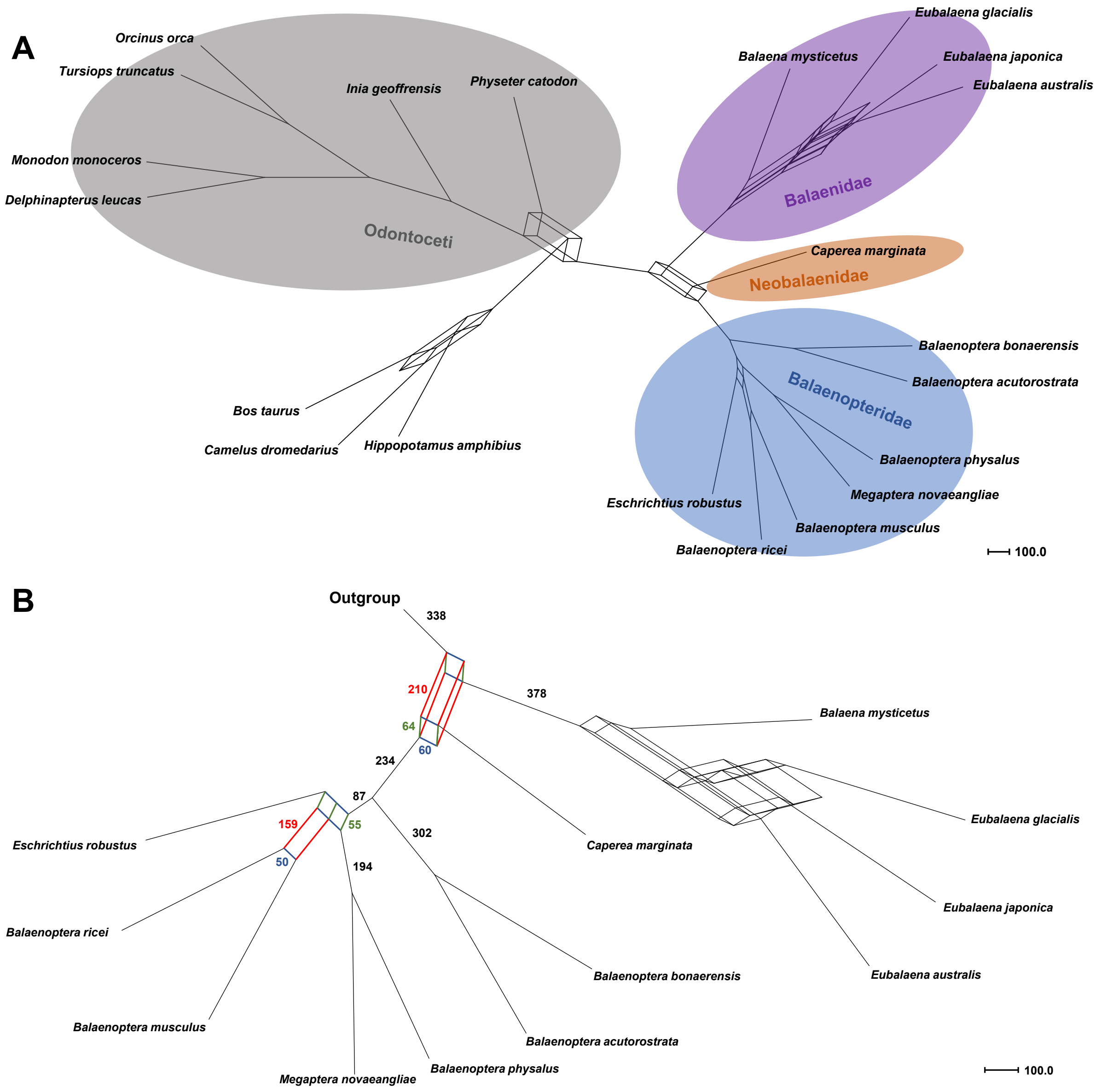

Fig. S5 Consensus network of Cetacea evolution based on single copy orthologous sequences. The network was conflated from gene trees that were constructed from 563 SCOS and a 12% threshold was used to depict conflicts. Extensive phylogenetic conflicts characterize the placement of the gray whale consistent with branch 4 of the main phylogenomic analysis. Additional conflicts were found at the placement of the pygmy right whale, at the base of the right whale divergence and at the branch separating toothed whales (*Odontoceti*) and baleen whales (*Mysticeti*), although they were less even.

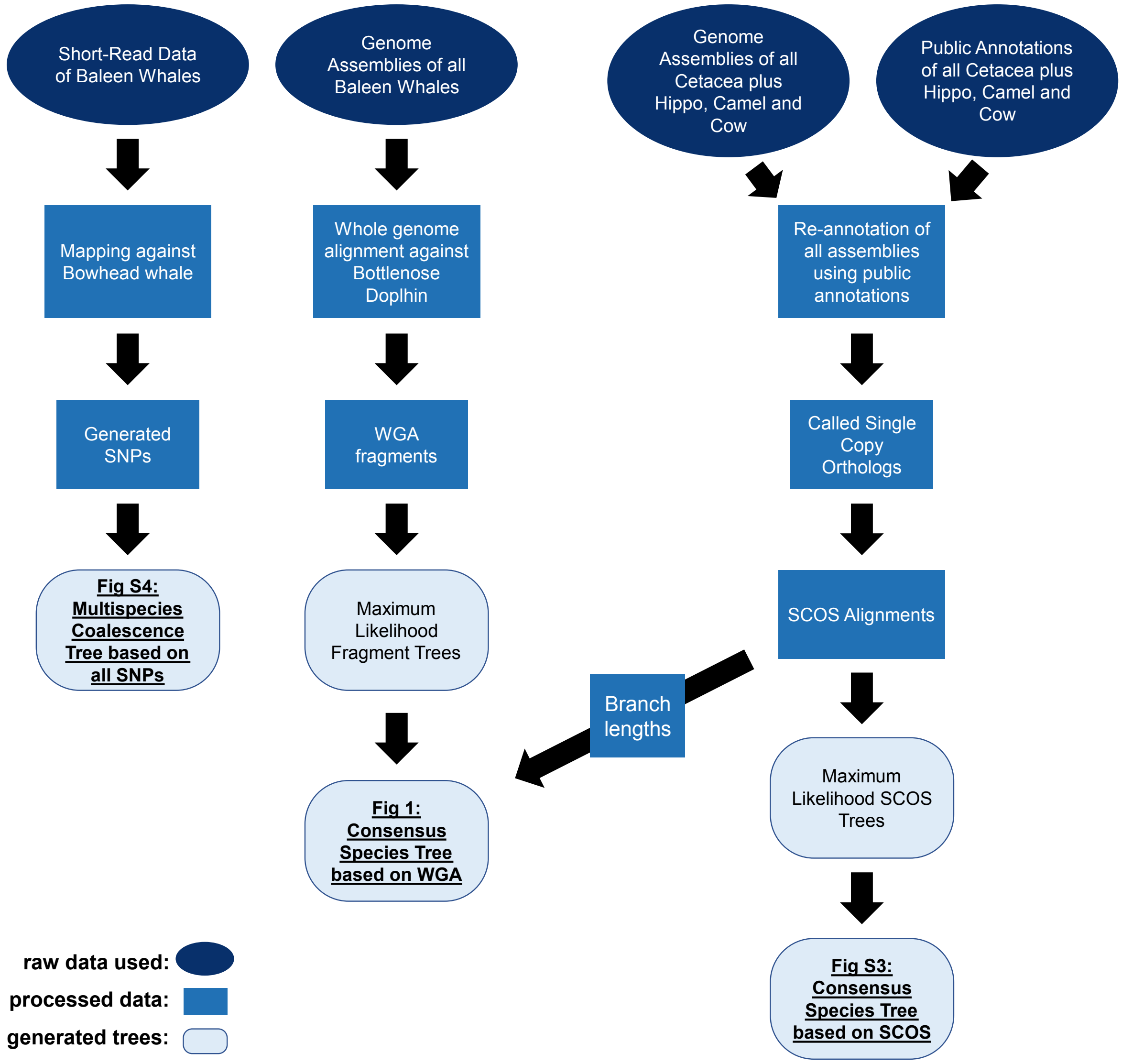
Short-Read Data of Baleen Whales
Genome Assemblies of all Baleen Whales
Genome Assemblies of all Cetacea plus Hippo, Camel and Cow
Public Annotations of all Cetacea plus Hippo, Camel and Cow
Mapping against Bowhead whale
Whole genome alignment against Bottlenose Doplhin
Re-annotation of all assemblies using public annotations
Generated SNPs
WGA fragments
Called Single Copy Orthologs
Fig S4: Multispecies Coalescence Tree based on all SNPs
Maximum Likelihood Fragment Trees
SCOS Alignments
Branch lengths
Fig 1: Consensus Species Tree based on WGA
Maximum Likelihood SCOS Trees
Fig S3: Consensus Species Tree based on SCOS
raw data used:
processed data:
generated trees:

Fig. S6 Pipeline depicting the process of generating all phylogenomic trees. It is differentiated between used raw input data (dark blue circles), processed intermediate data (blue rectangles) and resulting output trees (light blue rounded squares). Final trees presented in either the main manuscript or the supplement are underlined.

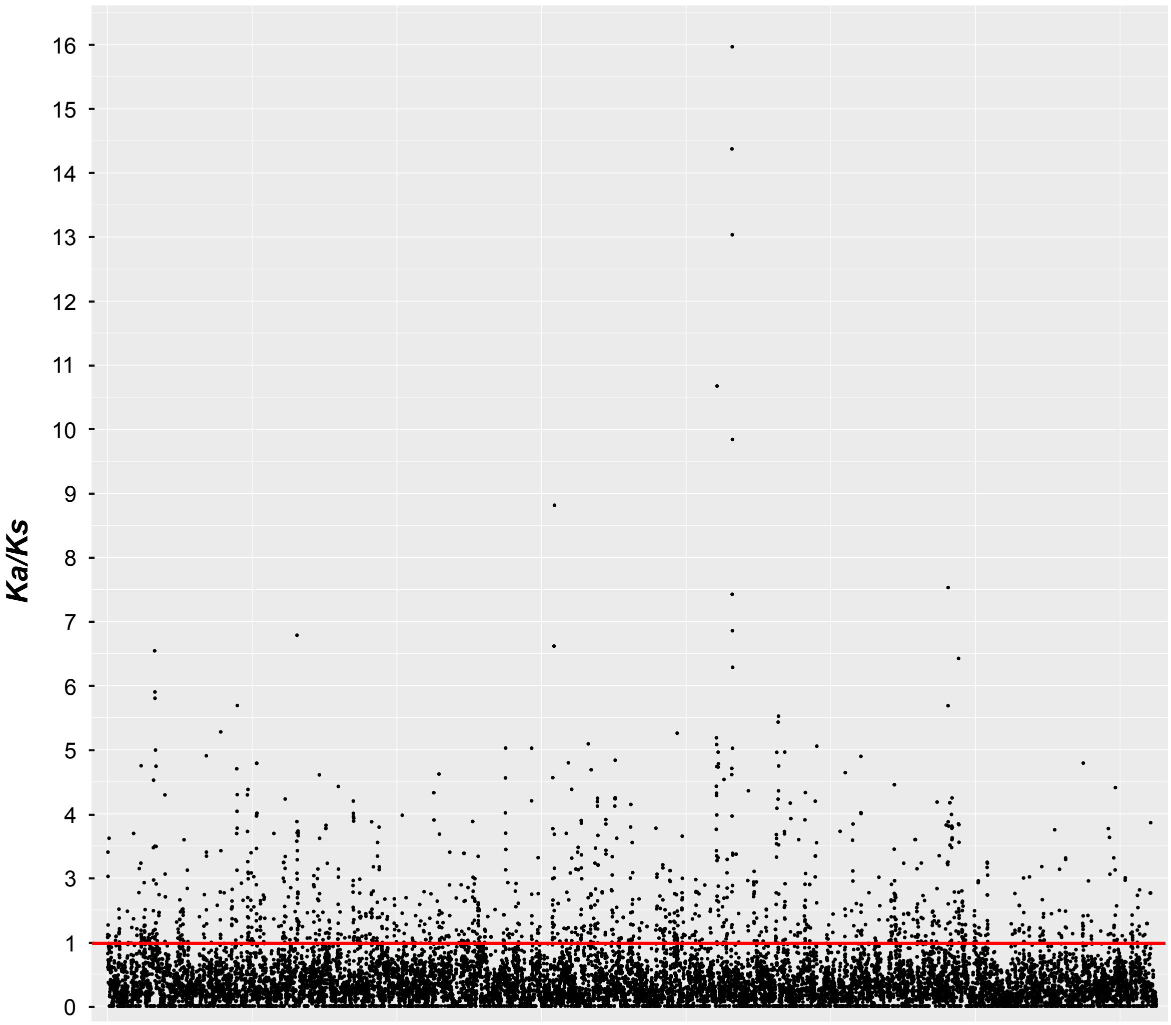

Fig. S7 Distribution of Ka/Ks values over all tested orthologs. Orthologs were inferred between all six baleen whales that resulted from the phylogenetic targeting analysis and the human genome GRCh38. Genes with Ka/Ks > 1 were considered as putative selected (red line) in downstream analyses.

Supplementary Tables

Table S1 Repeat content of the pygmy right whale assembly. Repeats were collected by REPEATMASKER v4.1 (www.repeatmasker.org) after modelling and identifying them with REPEATMODELER v2 (www.repeatmasker.org) using the Cetartiodactyla database from REPBASE (Jurka et al. 2005).

	Number of elements	**Length occupied (bp)**	**Percentage of sequence (%)**
Retroelements	2413632	865034774	34.39
SINEs:	690495	129626373	5.15
Penelope	66	12853	0.00
LINEs:	1354802	596935095	23.73
CRE/SLACS	0	0	0.00
L2/CR1/Rex	297003	65503588	2.60
R1/LOA/Jockey	0	0	0.00
R2/R4/NeSL	355	92064	0.00
RTE/Bov-B	10495	3115804	0.12
L1/CIN4	1046745	528167155	21.00
LTR elements:	368335	138473306	5.51
BEL/Pao	0	0	0.00
Ty1/Copia	286	1064306	0.04
Gypsy/DIRS1	14204	3919658	0.16
Retroviral	343447	130769605	5.20
DNA transposons:	369981	79396649	3.16
hobo-Activator	263101	52978695	2.11
Tc1-IS630-Pogo	97913	24838964	0.99
En-Spm	0	0	0.00
MuDR-IS905	0	0	0.00
PiggyBac	899	312728	0.01
Tourist/Harbinger	319	56939	0.00
Other transposons (Mirage,Pelement,Transib)	0	0	0.00
Others:			

Rolling-circles	1547	388897	0.02
Unclassified:	39979	6637640	0.26
Total interspersed repeats:		951069063	37.81
Small RNA:	325970	77630921	3.09
Satellites:	3700	1918182	0.08
Simple repeats:	8338	1139523	0.05
Low complexity:	0	0	0.00

Table S2 QuIBL results for all triplets resulting from combining rorquals species affecting the placement of the gray whale (excluding the two minke whale species). Tested were 1000 randomly selected WGA fragment trees.

triplet	outgr.	C1	C2	ILS-prop.	ILS+ Intro. prop.	num-trees	BIC (ILS + Intro)	BIC (only ILS)	ΔBIC	total prop. non-ILS
Bmus_Mnov_Bric	Bmus	0	52.736	0.987	0.013	77	-664.726	-605.133	-59.593	0.001
Bmus_Mnov_Bric	Bric	0	0.532	0.604	0.396	50	-497.555	-504.347	6.791	0.020
Bmus_Mnov_Bric	Mnov	0	1.537	0.010	0.990	873	-9325.190	-8838.495	-486.695	0.864
Bphy_Bmus_Bric	Bmus	0	7.415	0.939	0.061	88	-654.727	-649.388	-5.339	0.005
Bphy_Bmus_Bric	Bric	0	0.635	0.313	0.687	50	-513.413	-516.890	3.477	0.034
Bphy_Bmus_Bric	Bphy	0	1.560	0.011	0.989	862	-9208.359	-8723.705	-484.654	0.852
Bphy_Erob_Bric	Bphy	0	0.570	0.214	0.786	320	-3555.136	-3524.173	-30.963	0.251
Bphy_Erob_Bric	Erob	0	0.177	0.598	0.402	342	-3215.928	-3222.379	6.451	0.137
Bphy_Erob_Bric	Bric	0	1.315	0.249	0.751	338	-3845.348	-3797.613	-47.735	0.254
Bphy_Mnov_Bric	Bphy	0	0.706	0.487	0.513	31	-301.791	-307.331	5.541	0.016
Bphy_Mnov_Bric	Mnov	0	5.228	0.883	0.117	49	-380.309	-383.908	3.598	0.006
Bphy_Mnov_Bric	Bric	0	1.655	0.000	1.000	920	-9510.059	-8906.339	-603.720	0.920
Bphy_Erob_Bmus	Bphy	0	0.467	0.329	0.671	332	-3585.732	-3572.468	-13.264	0.223
Bphy_Erob_Bmus	Erob	0	0.835	0.208	0.792	327	-3595.573	-3554.387	-41.186	0.259
Bphy_Erob_Bmus	Bmus	0	0.592	0.108	0.892	341	-3774.736	-3717.975	-56.762	0.304
Bphy_Bmus_Mnov	Bphy	0	0.506	0.177	0.823	29	-275.500	-278.007	2.508	0.024
Bphy_Bmus_Mnov	Mnov	0	0.799	0.609	0.391	41	-408.184	-414.814	6.630	0.016
Bphy_Bmus_Mnov	Bmus	0	1.037	0.000	1.000	930	-9192.187	-8742.306	-449.881	0.930
Bphy_Erob_Mnov	Bphy	0	0.423	0.352	0.648	49	-457.953	-462.624	4.671	0.032
Bphy_Erob_Mnov	Mnov	0	9.184	0.941	0.059	41	-380.572	-377.169	-3.403	0.002
Bphy_Erob_Mnov	Erob	0	0.844	0.001	0.999	910	-8847.166	-8491.310	-355.856	0.909
Erob_Bmus_Bric	Bmus	0	0.270	0.773	0.227	73	-661.116	-669.611	8.495	0.017
Erob_Bmus_Bric	Bric	0	0.278	0.238	0.762	74	-678.588	-681.750	3.161	0.056
Erob_Bmus_Bric	Erob	0	1.755	0.138	0.862	853	-8783.866	-8510.811	-273.055	0.736
Erob_Mnov_Bric	Erob	0	0.264	0.476	0.524	339	-3312.733	-3313.683	0.950	0.178
Erob_Mnov_Bric	Mnov	0	0.471	0.277	0.723	328	-3545.221	-3527.425	-17.796	0.237
Erob_Mnov_Bric	Bric	0	0.557	0.148	0.852	333	-3655.329	-3611.483	-43.847	0.284
Erob_Bmus_Mnov	Erob	0	0.858	0.188	0.812	77	-3620.386	-3573.493	-46.893	0.266
Erob_Bmus_Mnov	Mnov	0	0.772	0.582	0.418	50	-3641.136	-3645.268	4.132	0.141
Erob_Bmus_Mnov	Bmus	0	0.550	0.118	0.882	873	-3660.379	-3612.023	-48.355	0.294

triplet: The three-taxon subset considered. Species abbreviations separated by underscores
outgr.: Species inferred to be the outgroup out of the considered triplet.
C(n): Inferred species tree branch length for (1) the ILS-only model and (2) the ILS+introgression model. The ILS model is forced to be 0, as all lineages must be in the same population.
ILS-prop.: inferred proportion of the trees that account for the ILS-only model
ILS+Intro. prop.: inferred proportion of the trees that account for the ILS+introgression model.
num-trees: number of trees in the considered topology
BIC: Value resulted from a Bayesian information criterion test
BIC (n): raw BIC values for one of both models
ΔBIC: difference in BIC value between the models. ΔBIC < -10 was set as a cutoff to decide between the ILS-only model (ΔBIC > -10) and the ILS+introgression model (ΔBIC < -10)
total prop.: total proportion of trees that support the ILS+introgression model calculated as “ILS+Intro. prop * (num-trees / total trees in sample)”.

Table S3 Calibration points used in the date phylogeny.

Defined Node	Fossil record (species)	Suggested age (in Mya)	Literature
Cetartiodactyla	*Diacodexis ilicis*	57.3 - 72.2	Gingerich 1989
Cetancodonta	*Himalayacetus subathuensis*	52.9 - 58.3	Bajpai and Gingerich 1998
Cetacea	*Mystacodon selenensis*	36.1 - 39.6	Lambert et al. 2017
Mysticeti	*Balaenella brachyrhynus*	21.1 - 26.8	Bisconti 2005
Odontoceti	*Arktocara yakataga*	25.1 - 30.8	Boersma and Pyenson 2016

Table S4 Body mass data used for phylogenetic targeting. We collected estimates for mean body-mass (Kg), mean body-length (m) and longevity (years). Eventually, longevity was excluded from the phylogenetic targeting analysis due to the lack of information available.

Species	Length (m)	Mass (Kg)	Longevity	Literature
Caperea marginata	6	3430	?	Budylenko et al. 1973
Balaena mysticetus	19	80000	200	Georg et al. 1999
Eubalaena australis	14	35000	?	Hamilton et al. 1998, Fortune et al. 2021
Eubalaena glacialis	11	35000	70	Christiansen et al. 2019
Eubalaena japonica	16	60000	?	Lockyer 1976
Balaenoptera bonaerensis	9	6800	?	Konishi 2006
Balaenoptera acutorostrata	8,5	5000	50	Markussen et al. 1992, Horwood 1989
Balaenoptera musculus	23	100000	90	Ruud 1956, Gilpatrick and Perryman, Sears and Perrin 2008
Balaenoptera ricei	12	13000	?	Tershy 1992, Rosel et al. 2021
Eschrichtius robustus	14	35000	77	Rice and Wolman 1971, Swartz 2018
Megaptera novaeangliae	14	40000	95	Chittleborough 1959, Jefferson et al. 2015, Clapham and Mead 1999
Balaenoptera physalus	20	45000	90	Lokyer and Waters 1986, Aguilar and Garcia-Vernet 2018

Table S5 Maximal pairs inferred from the phylogenetic targeting analysis. Pairs were inferred using the PhyloTargeting Webserver (https://phylotargeting.nunn-lab.org/index.html) by providing body length and mass data (Supplement Table S4) as well as the phylogenetic tree depicted in Fig. 1. Tursiops truncates was used as an outgroup. Eventually, we did not use the pair of right whales due to uncertainty in the data as well as having the lowest standardized summed score.

Species 1	Species 2	Raw difference (body mass)	Score (body mass)	Raw difference (body length)	Score (body length)	Sum of branch lengths	No. of branches	Summed score (standard.)
Cmar	**Bmys**	**76.57**	**0.793**	**13**	**0.765**	**0.081**	**4**	**1.558**
Egla	**Ejap**	**25**	**0.259**	**4.5**	**0.265**	**0.017**	**2**	**0.524**
Bacu	**Bphy**	**40**	**0.414**	**11.5**	**0.676**	**0.034**	**6**	**1.091**
Bric	**Bmus**	**87**	**0.901**	**11**	**0.647**	**0.019**	**2**	**1.548**

Table S6 Used data featured in this study including assemblies, short read archives and proteomes from other *Cetacea* or *Cetartiodactyla*. Provided are information for the scientific and common name, for the database, the respective ID, source publication or project and usage in this study. A list of aberrations can be found below.

Assemblies

Species	Common Name	Database	ID	Literature/Consortium	Usage
Balaena mysticetus	bowhead whale	Bowhead whale genome resource	-	Keane et al. 2015	WGA,SCOS,SNP reference,cancer analyses reference
Eubalaena australis	Southern right whale	DNA Zoo	-	DNA Zoo	WGA,SCOS
Eubalaena glacialis	North Atlantic right whale	DNA Zoo	-	DNA Zoo	WGA,SCOS
Eubalaena japonica	North Pacific right whale	NCBI	GCA_004363455.1	Zoonomia Consortium	WGA,SCOS
Balaenoptera bonaerensis	Antarctic minke whale	NCBI	GCA_000978805.1	Kishida et al. 2015	WGA,SCOS

Balaenoptera acutorostrata	minke whale	NCBI	GCA_000493695.1	Yim et al. 2014	WGA,SCOS,cancer analyses reference
Balaenoptera musculus	blue whale	CNGBdb	CNA0007254	Yuan et al. 2021	WGA,SCOS,cancer analyses reference
Balaenoptera ricei	rice whale	DNA Zoo	-	DNA Zoo	WGA,SCOS
Eschrichtius robustus	gray whale	NCBI	GCA_004363415.1	Zoonomia Consortium	WGA,SCOS
Megaptera novaeangliae	humpback whale	NCBI	GCA_004329385.1	Tollis et al. 2019	WGA,SCOS,cancer analyses reference
Balaenoptera physalus	fin whale	NCBI	GCA_023338255.1	Wolf et al. 2022	WGA,SCOS
Bos taurus	cattle	NCBI	GCA_002263795.3 ARS-UCD1.3	USDA ARS	SCOS
Camelus dromedarius	Arabian camel	NCBI	GCA_000803125.3 CamDro3	Elbers et al. 2019	SCOS
Hippopotamus amphibius	hippopotamus	NCBI	GCA_023065835.1 ASM2306583v1	Northwestern Polytechnology University	SCOS
Physeter catodon	sperm whale	NCBI	GCA_002837175.2 ASM283717v2	Fan et al. 2018	SCOS
Inia geoffrensis	boutu	NCBI	GCA_004363515.1 IniGeo_v1_BIUU	Broad Institute	SCOS
Tursiops truncatus	bottlenose dolphin	NCBI	GCA_011762595.1	VGP	WGA reference,SCOS
Orcinus orca	killer whale	NCBI	GCA_000331955.2 Oorc_1.1	Foote et al. 2015	SCOS
Delphinapterus leucas	beluga whale	NCBI	GCA_002288925.3 ASM228892v3	Jones et al. 2017	SCOS

Monodon monoceros	narwhal	NCBI	GCA_005125345.1	Westbury et al. 2019	SCOS

SRA

Species	Common Name	Database	ID	Literature/Consortium	Usage
Balaena mysticetus	bowhead whale	Bowhead whale genome resource	-	Keane et al. 2015	SNP,cancer analyses
Eubalaena glacialis	North Atlantic right whale	NCBI	SRR11097130	DNA Zoo	SNP
Balaenoptera acutorostrata	minke whale	NCBI	SRR924087	Yim et al. 2014	SNP,cancer analyses
Megaptera novaeangliae	humpback whale	NCBI	SRP175048	Tollis et al. 2019	SNP,cancer analyses
Balaenoptera physalus	fin whale	NCBI	SRP325690	Wolf et al. 2022	SNP, cancer analyses
Eschrichtius robustus	gray whale	NCBI	SRP108933	Arnason et al. 2018	SNP
Balaenoptera musculus	blue whale	NCBI	SRP108933	Arnason et al. 2018	SNP, cancer analyses
Balaenoptera borealis	sei whale	NCBI	SRP108933	Arnason et al. 2018	SNP

Proteoms

Species	**Common Name**	**database**	**ID**	**Literature/Consortium**	**Usage**
Balaena mysticetus	bowhead whale	Bowhead whale genome resource	-	Keane et al. 2015	Cmar annotation,SCOS, cancer analyses
Balaenoptera acutorostrata	minke whale	NCBI	GCF_000493695	Yim et al. 2014	Cmar annotation,SCOS, cancer analyses
Balaenoptera physalus	fin whale	NCBI	GCA_023338255.1	Wolf et al. 2022	Cmar annotation,SCOS, cancer analyses
Bos taurus	cattle	NCBI	GCA_002263795.3 ARS-UCD1.3	USDA ARS	Cmar annotation,SCOS, cancer analyses
Camelus dromedarius	Arabian camel	NCBI	GCA_000803125.3 CamDro3	Elbers et al. 2019	Cmar annotation,SCOS, cancer analyses
Delphinapterus leucas	beluga whale	NCBI	GCA_002288925.3 ASM228892v3	Jones et al. 2017	Cmar annotation,SCOS, cancer analyses
Monodon monoceros	narwhal	NCBI	GCA_005125345.1	Westbury et al. 2019	Cmar annotation,SCOS, cancer analyses
Orcinus orca	killer whale	NCBI	GCA_000331955.2 Oorc_1.1	Foote et al. 2015	Cmar annotation,SCOS, cancer analyses
Physeter catodon	sperm whale	NCBI	GCA_002837175.2 ASM283717v2	Fan et al. 2018	Cmar annotation,SCOS, cancer analyses
Tursiops truncatus	bottlenose dolphin	NCBI	GCA_011762595.1	VGP	Cmar annotation,SCOS, cancer analyses

NCBI: National Center for Biotechnology Information
CNGBdb: China National GeneBank DataBase
WGA: Whole-genome Alignment approach (Fig 1, main manuscript)
SCOS: Single Copy Orthologous Sequence approach (Fig. S2)
SNP: Single Nucleotide Polymorphism approach (Fig. S3)
VGP: Vertebrate Genome Project
Cmar: *Caperea marginata*

Supplementary References

Aguilar A, García-Vernet R. Fin whale: Balaenoptera physalus. In: Würsig BG, Thewissen JGM, Kovacs KM, eds. Encyclopedia of marine mammals. Amsterdam: Academic Press; 2017. p. 368–371.

Árnason Ú, Lammers F, Kumar V, Nilsson MA, Janke A. Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. Sci Adv. 2018;4(4):eaap9873. doi:10.1126/sciadv.aap9873.

Bajpai S, Gingerich PD. A new Eocene archaeocete (Mammalia, Cetacea) from India and the time of origin of whales. Proc Natl Acad Sci U S A. 1998;95(26):15464–8. doi:10.1073/pnas.95.26.15464.

Bisconti M. SKULL MORPHOLOGY AND PHYLOGENETIC RELATIONSHIPS OF A NEW DIMINUTIVE BALAENID FROM THE LOWER PLIOCENE OF BELGIUM. Palaeontology. 2005;48(4):793–816. doi:10.1111/j.1475-4983.2005.00488.x.

Boersma AT, Pyenson ND. Arktocara yakataga, a new fossil odontocete (Mammalia, Cetacea) from the Oligocene of Alaska and the antiquity of Platanistoidea. PeerJ. 2016;4:e2321. doi:10.7717/peerj.2321.

Budylenko GA, Panfilov BG, Pakhomova AA, Sazhinov EG. New data on pygmy right whales Neobalaena marginata (Gray, 1848). Trudy Atlanticheskii Nauchno-Issledovatel'skii Institut Rybnogo Khozyaistva I Okeanografii. 1973;51:122–32.

Chittleborough RG. Determination of age in the humpback whale, Megaptera nodosa (Bonnaterre). Marine and Freshwater Research. 1959;10(2):125–43.

Christiansen F, Sironi M, Moore MJ, Di Martino M, Ricciardi M, Warick HA, et al. Estimating body mass of free-living whales using aerial photogrammetry and 3D volumetrics. Methods Ecol Evol. 2019;10(12):2034–44. doi:10.1111/2041-210X.13298.

Clapham PJ, Mead JG. Megaptera novaeangliae. Mammalian Species. 1999(604):1. doi:10.2307/3504352.

Elbers JP, Rogers MF, Perelman PL, Proskuryakova AA, Serdyukova NA, Johnson WE, et al. Improving Illumina assemblies with Hi-C and long reads: An example with the North African dromedary. Mol Ecol Resour. 2019;19(4):1015–26. doi:10.1111/1755-0998.13020.

Fan G, Zhang Y, Liu X, Wang J, Sun Z, Sun S, et al. The first chromosome-level genome for a marine mammal as a resource to study ecology and evolution. Mol Ecol Resour. 2019;19(4):944–56. doi:10.1111/1755-0998.13003.

Foote AD, Liu Y, Thomas GWC, Vinař T, Alföldi J, Deng J, et al. Convergent evolution of the genomes of marine mammals. Nat Genet. 2015;47(3):272–5. doi:10.1038/ng.3198.

Fortune SME, Moore MJ, Perryman WL, Trites AW. Body growth of North Atlantic right whales (Eubalaena glacialis) revisited. Marine Mammal Science. 2021;37(2):433–47. doi:10.1111/mms.12753.

George JC, Bada J, Zeh J, Scott L, Brown SE, O'Hara T, Suydam R. Age and growth estimates of bowhead whales (Balaena mysticetus) via aspartic acid racemization. Can. J. Zool. 1999;77(4):571–80. doi:10.1139/z99-015.

Gilpatrick JW, Perryman WL. Geographic variation in external morphology of North Pacific and Southern Hemisphere blue whales (Balaenoptera musculus). Journal of Cetacean Research and Management. 2008;10(1):9–21.

Gingerich PD. New Earliest Wasatchian Mammalian Fauna from the Eocene of Northwestern Wyoming: Composition and Diversity in a Rarely Sampled High-Floodplain Assemblage. 1989.

Hamilton PK, Knowlton AR, Marx MK, Kraus SD. Age structure and longevity in North Atlantic right whales Eubalaena glacialis and their relation to reproduction. Mar. Ecol. Prog. Ser. 1998;171:285–92. doi:10.3354/meps171285.

Horwood J. Biology and exploitation of the minke whale. Boca Raton, Fla.: CRC Press; 1990.

Jefferson TA, Pitman RL, Webber MA, editors. Marine mammals of the world: A comprehensive guide to their identification. 2nd ed. Amsterdam: Academic Press; 2015.

Jones SJM, Taylor GA, Chan S, Warren RL, Hammond SA, Bilobram S, et al. The Genome of the Beluga Whale (Delphinapterus leucas). Genes (Basel) 2017. doi:10.3390/genes8120378.

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1-4):462–7. doi:10.1159/000084979.

Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, et al. Insights into the evolution of longevity from the bowhead whale genome. Cell Rep. 2015;10(1):112–22. doi:10.1016/j.celrep.2014.12.008.

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

Lambert O, Martínez-Cáceres M, Bianucci G, Di Celma C, Salas-Gismondi R, Steurbaut E, et al. Earliest Mysticete from the Late Eocene of Peru Sheds New Light on the Origin of Baleen Whales. Curr Biol. 2017;27(10):1535-1541.e2. doi:10.1016/j.cub.2017.04.026.

Lockyer C. Body weights of some species of large whales. ICES Journal of Marine Science. 1976;36(3):259–73. doi:10.1093/icesjms/36.3.259.

Lockyer C, Waters T. WEIGHTS AND ANATOMICAL MEASUREMENTS OF NORTHEASTERN ATLANTIC FIN (BALAENOPTERA PHYSALUS, LINNAEUS) AND SEI (B. BOREALIS, LESSON) WHALES. Marine Mammal Science. 1986;2(3):169–85. doi:10.1111/j.1748-7692.1986.tb00039.x.

Markussen NH, Ryg M, Lydersen C. Food consumption of the NE Atlantic minke whale (Balaenoptera acutorostrata) population estimated with a simulation model. ICES Journal of Marine Science. 1992;49(3):317–23. doi:10.1093/icesjms/49.3.317.

Rice DW, Wolman AA. The life history and ecology of the gray whale (Eschrichtius robustus). American Society of Mammalogist; 1971.

Rosel PE, Wilcox LA, Yamada TK, Mullin KD. A new species of baleen whale (Balaenoptera) from the Gulf of Mexico, with a review of its geographic distribution. Marine Mammal Science. 2021;37(2):577–610. doi:10.1111/mms.12776.

Ruud JT. The Blue Whale. Scientific American. 1956;195(6):46–51.

Sears R, Perrin WF. Blue Whale: Balaenoptera musculus. In: Perrin WF, Würsig BG, Thewissen JGM, eds. Encyclopedia of marine mammals. 2nd ed. Amsterdam, Boston, Mass.: Elsevier/Academic Press; 2009. p. 120–124.

Swartz SL. Gray Whale: Eschrichtius robustus. In: Würsig BG, Thewissen JGM, Kovacs KM, eds. Encyclopedia of marine mammals. Amsterdam: Academic Press; 2017. p. 422–428.

Tershy BR. Body Size, Diet, Habitat Use, and Social Behavior of Balaenoptera Whales in the Gulf of California. Journal of Mammalogy. 1992;73(3):477–86. doi:10.2307/1382013.

Tollis M, Robbins J, Webb AE, Kuderna LFK, Caulin AF, Garcia JD, et al. Return to the Sea, Get Huge, Beat Cancer: An Analysis of Cetacean Genomes Including an Assembly for the Humpback Whale (Megaptera novaeangliae). Mol Biol Evol. 2019;36(8):1746–63. doi:10.1093/molbev/msz099.

Westbury MV, Petersen B, Garde E, Heide-Jørgensen MP, Lorenzen ED. Narwhal Genome Reveals Long-Term Low Genetic Diversity despite Current Large Abundance Size. iScience. 2019;15:592–9. doi:10.1016/j.isci.2019.03.023.

Wolf M, Jong M de, Halldórsson SD, Árnason Ú, Janke A. Genomic Impact of Whaling in North Atlantic Fin Whales. Mol Biol Evol 2022. doi:10.1093/molbev/msac094.

Yim H-S, Cho YS, Guang X, Kang SG, Jeong J-Y, Cha S-S, et al. Minke whale genome and aquatic adaptation in cetaceans. Nat Genet. 2014;46(1):88–92. doi:10.1038/ng.2835.

Yuan Y, Zhang Y, Zhang P, Liu C, Wang J, Gao H, et al. Comparative genomics provides insights into the aquatic adaptations of mammals. Proc Natl Acad Sci U S A 2021. doi:10.1073/pnas.2106080118.

6.3 Manuscript 1:

Genomes of northern hemisphere and pygmy blue whales suggest isolated subspecies.

Magnus Wolf, Menno J. de Jong, Aimee R. Lang, Axel Janke

Submitted as a research article to: *Molecular Ecology* (24 April 2023)

Declaration of author contributions to the publication

Manuscript: Genomes of northern hemisphere and pygmy blue whales suggest isolated subspecies.

Status: submitted

Name of the journal: Molecular Ecology

Contributing Authors:

Magnus Wolf (MW), Menno J. de Jong (MdJ), Axel Janke (AJ)

Contributions of the doctoral candidate and his co-authors:

1.) Concept and design		
MW	80%	
AJ	20%	
2.) Data analysis and figure compilation		
MW	90%	Compilation of all datasets and figures
MdJ	10%	Contribution to diversity and divergence statistics
3.) Interpretation of data		
MW	80%	Analysis and interpretation of all data
AJ	20%	Summary interpretation
4.) Drafting manuscript		
MW	75%	
MdJ	5%	
AJ	20%	
5.) Other works		
MW	100%	Sample organisation

I hereby certify that the information above is correct.

Date and place

Signature doctoral candidate

Date and place

Signature supervisor

MOLECULAR ECOLOGY

Genomes of northern hemisphere and pygmy blue whales suggest isolated subspecies.

Journal:	*Molecular Ecology*
Manuscript ID	Draft
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Wolf, Magnus; Senckenberg Gesellschaft fur Naturforschung, Biodiversity and Climate Research Centre; Goethe University Frankfurt, Institute for Ecology, Evolution & Diversity de Jonog, Menno; Senckenberg Gesellschaft fur Naturforschung Janke, Axel; LOEWE-TBG; Senckenberg Gesellschaft fur Naturforschung; Goethe University Frankfurt, Institue for Ecology, Evolution & Diversity
Keywords:	subspecies, blue whales, population genomics, conservation genomics, genetic diversity, runs of homozygosity

Genomes of northern hemisphere and pygmy blue whales suggest isolated subspecies.

Magnus Wolf[1,2], Menno J. de Jong[1], Axel Janke[1,2,3]

[1] Senckenberg Biodiversity and Climate Research Centre (BiK-F), Georg-Voigt-Strasse 14-16, Frankfurt am Main, Germany

[2] Institute for Ecology, Evolution and Diversity, Goethe University, Max-von-Laue-Strasse. 9, Frankfurt am Main, Germany

[3] LOEWE-Centre for Translational Biodiversity Genomics (TBG), Senckenberg Nature Research Society, Georg-Voigt-Straße 14-16, Frankfurt am Main, Germany

Author for correspondence:
Magnus Wolf
Phone: +49 69 754218-28 | Email: Magnus.Wolf@senckenberg.de

Name	e-mail	ORCID
Magnus Wolf	magnus.wolf@senckenberg.de	0000-0001-9212-9861
Menno J. de Jong	menno.de-jong@senckenberg.de	0000-0003-2131-9048
Axel Janke	axel.janke@senckenberg.de	0000-0002-9394-1904

Abstract

The blue whale is an endangered and globally distributed species of baleen whale with multiple described subspecies assignments like the morphologically and molecularly distinct pygmy blue whale. Despite considerable acoustic differences between Atlantic and Pacific populations, both are currently regarded as a single subspecies that can be considered a remnant of other subspecies definitions in the southern hemisphere. To determine the degree of isolation among northern hemisphere blue whales, we sequenced 14 North-Pacific blue whales and 6 Indo-Australian pygmy blue whale genomes and mitogenomes. Together with publicly available samples of North-Atlantic blue whales, we studied different aspects of the genetic differentiation and genetic exchange among northern-hemisphere populations in the context of the well described pygmy blue whale subspecies. Population genomic analyses revealed highly differentiated clusters and limited exchange between all three populations, supporting their high degree of isolation, while the genomic and mitogenomic differences in all blue whales including the pygmy blue whale subspecies are low. Given that the pygmy blue whale is an already established subspecies and supported by distinct acoustic differences, we propose to treat the two northern hemisphere populations equally as subspecies and suggest names for the North-Atlantic and North-Pacific blue whale. Furthermore, we provide a first molecular viability assessment of all three blue whale populations that showcases the generally high genomic diversity among blue whales but also a lack of rare alleles and increased inbreeding suggestive for the anthropogenic impact on the genotypes of blue whales. With this, we provide much needed genomic insights into northern hemisphere blue whales to guide further conservation efforts of this threatened and iconic species.

Keywords

subspecies, blue whales, population genomics, conservation genomics, genetic diversity, runs of homozygosity.

Introduction

With maximum body sizes exceeding 30 m and mean body sizes between 21-26 m (Branch, T. A., Abubaker, E. M. N., Mkango, S., & Butterworth, D. S., 2007; McClain et al., 2015), the blue whale (*Balaenoptera musculus*) is the largest baleen whale (*Mysticeti*) and the largest species known to have ever existed on earth (Sears & Perrin, 2009). The blue whale can be found in all major oceans and its distribution is only restricted by oligotrophic central-oceanic regions that lack impactful ocean dynamics like upwelling and frontal meandering (Branch, T. A. et al., 2007). Most individuals are known to participate in seasonal migrations between productive feeding grounds and warmer or safer breeding grounds that require them to travel large distances (Burtenshaw et al., 2004; Double et al., 2014) In some cases, individuals have been recorded to travel distances of up to 8,000 kilometers during their lifetime (Hucke-Gaete et al., 2018), showcasing their ability for trans-oceanic migrations.

Multiple blue whale subspecies with different distributions, morphologies, acoustic, and molecular patterns have been proposed over the years (Branch, T. A. et al., 2007; Leduc et al., 2007; Sears & Perrin, 2009, 2009). The most distinctive subspecies, the pygmy blue whale (*B. m. brevicauda*) is the smallest subspecies and typically found near Australian waters in the eastern Indian Ocean and western tropical Pacific (Branch, T. A. et al., 2007). They were first differentiated from the southern, Antarctic blue whale (*B. m. intermedia*) by studies on body size and other morphological features (Ichihara, 1966), which was later supported by acoustic and eventually genetic evidence (Leduc et al., 2007; Ljunglad, Clark, & Shimada, 1998). There is also a general distinction between whales from the southern hemisphere (*B. m. intermedia*) and whales from the northern hemisphere (*B. m. musculus*) based on body size differences and the expectation of limited gene flow across equatorial areas (Sears & Perrin, 2009). While there is an ongoing discussion of whether south-eastern Pacific blue whales, distributed alongside the coast of Chile, form a subspecies distinct from *B. m. brevicauda* and *B. m. intermedia* (Branch, T. A. et al., 2007; Leduc et al., 2007; Pastene, Acevedo, & Branch, 2019), blue whales from the northern hemisphere were usually not divided any further, despite recognized differences in acoustics (McDonald, Mesnick, & Hildebrand, 2006; Mellinger & Clark, 2003) and an expectable limited gene flow given the separation by the continental landmasses.

Like all baleen whales, the number of blue whales were reduced dramatically by 20th century commercial hunting (Tønnessen & Johnsen, 1982). After whaling restrictions were established by the International Whaling Commission (IWC) in 1986, the species appears to have recovered in numbers in some parts of its range, although it is still classified as endangered under the IUCN red list (Cooke, 2018). From a genetic point of view, it can be assumed that this bottleneck reduced the genetic diversity within blue whale populations, increased the likelihood of inbreeding, and altered gene flow between subpopulations and other species. First genome-wide estimates of their genetic diversity, however, revealed contradicting high levels of genome-wide heterozygosity (Bukhman et al., 2022; Jossey et al., 2021; Wolf, Jong, Halldórsson, Árnason, & Janke, 2022). Additionally, patterns of inbreeding, measured by the means of runs of homozygosity (ROH), indicated higher levels of inbreeding in the blue whale compared to other whale species, although these assessments were based on single genomes (Wolf et al., 2022). This picture is complicated by evidence of gene flow between blue whales and fin whales, *Balaenoptera physalus* (Jossey et al., 2021; Pampoulie et al., 2021) and among different blue whale subspecies (Attard et al., 2012). In the latter case, neither the extent of gene flow nor its impact on the genetic diversity of certain blue whale subspecies is known.

To guide conservation efforts of the species, it is crucial to know the extent of genetic exchange between different species, subspecies, and populations to characterize and delineate precise conservation units. While genetic exchange between the blue whale and the fin whale as well as between the Antarctic blue whale and the pygmy blue whale were addressed in previous studies (Attard et al., 2012; Jossey et al., 2021; Pampoulie et al., 2021), there is no information about admixture in the northern-hemisphere blue whale populations (*B. m. musculus*). Furthermore, conservation genomic analyses to assess genetic diversity, inbreeding, and mutational load can provide valuable information on the genetic viability of a population and could help to precisely channel further conservation efforts.

In this study, we estimate the genomic isolation among the two northern hemisphere blue whale populations from the North Atlantic and North Pacific and compare its extent to the well-defined pygmy blue whale subspecies. We sequenced the genomes from 20 blue whale specimens gathered from the North Pacific blue whale and West-Australian pygmy blue whale populations and analyze the data together with 12 publicly available genomes of other blue whales, including those of the North-Atlantic blue whales (Jossey et al., 2021). This sampling is further complemented by three genomes of the closely related sei whale (*Balaenoptera borealis*), of which one is sequenced in this study. This allows us to characterize the extent of gene flow and hence admixture between the populations, to determine the degree of genetic distance and to estimate their demographic past. Furthermore, the genetic viability of these populations is studied by the means of genome-wide heterozygosity and an inbreeding analysis based on ROH. With this, we provide a first evaluation of the differentiation between Pacific and Atlantic blue whales and provide valuable data to guide further conservation efforts of this endangered species.

Materials and Methods

Sampling, DNA, library, sequencing

Blue whale DNA samples were provided by the MMaSTR collection hosted by the NOAA Southwest Fisheries Science Center (SWFSC) and respective metadata are provided in the Supplement (Table S1). Hence, sampling procedures, sample handling and DNA extraction methods may vary but all samples were stored at a minimum of -20°C degrees Celsius and most DNA extractions were performed using a salting-out protocol (Miller, Dykes, & Polesky, 1988).

The DNA from the sei whale specimen was isolated from a tissue culture established by Úlfur Árnason from an Icelandic individual in the 1980s. The fibroblast-like cells were grown under standard conditions in rich DMEM medium supplemented with 5% FCS. After trypsinization and resuspension in standard homogenization buffer, the DNA was purified using standard phenol/chloroform method (Sambrook & Russell, 2006).

All Illumina paired-end libraries were prepared by Novogene, Cambridge, United Kingdom using the NEBNext DNA Library Prep kit with a read length of 150 base pairs (bp) and an insert size of 350 bp. Illumina sequencing was performed on a NovaSeq 6000 platform targeting ~25x coverage per individual.

Mapping, variant calling

Generated short read datasets from Pacific blue whales, West-Australian blue whales and of the single sei whale individual were combined with publicly available short read datasets from other blue and sei whales (Table S2). A comprehensive pipeline used to process the data and perform many of the here presented downstream analyses can be found on GitHub: mag-wolf/RESEQ-to-Popanalyses/.

Short read data were trimmed for quality and adapter sequences using FASTP V0.23.2 (Chen, Zhou, Chen, & Gu, 2018) with the options "-g -3 -l 40 -y -c -p". Trimmed reads were mapped to a repeat-masked, high-quality blue whale reference genome, constructed by the authors of the Vertebrate Genome Project (Bukhman et al., 2022). Mapping was performed using BWA MEM v0.7.17-r1188 (http://bio-bwa.sourceforge.net) and SAMtools v1.9 sort (Danecek et al., 2021) using default settings. Potential duplicates were removed and read-groups were added using the PICARD v2.21.2-0 toolkit (https://broadinstitute.github.io/picard/). Genotype-calling of autosomes, sex-chromosomes and mitochondrial genomes was done on all individuals combined and on each individual independently using BCFTOOLS v1.12 MPILEUP and BCFTOOLS v1.12 CALL (DANECEK ET AL., 2021) with the respective "-m" or "-c" flag and minimum mapping- and base-quality cutoffs of 20 and 13, respectively. For the sex-chromosome and mitochondrial genome data, BCFTOOLS CALL was run with the "--ploidy 1" flag to account for the haploidy. All inferred sites were further filtered by excluding sites with divergent read coverage (>3-fold and <0.3-fold of the expected individual mean coverage) and sites with more than 25% missing data using the BCFtools filter function. The combined genotype set was further processed by removing multivariate and monomorphic sites to retrieve single nucleotide polymorphisms (SNPs) with VCFTOOLS v0.1.16 (DANECEK ET AL., 2011) and by thinning SNPs to account for the effects of linkage disequilibrium using BCFTOOLS THINN function with a window size of 1000 bp. To receive variances of the mitochondrial control region (mtMarker), we extracted respective 403 bp long area from the mitogenomic vcf file using BCFTOOLS VIEW "-r" (See Rosel, Dizon, & Heyning, 1994). Eventually, a comprehensive table of sequencing, mapping and variant calling statistics can be found in Table S3.

Population structure and gene flow

Population-genetic analyses were performed using the R based tool collection SAMBAR v1.09 (Jong, Jong, Hoelzel, & Janke, 2021). Using SambaR's "findstructure()" function, a principal coordinate analysis (PCoA) based on the APE-5.3 package (Paradis & Schliep, 2019) as well as a an admixture analysis based on the LEA-2.4.0 package (Frichot & François, 2015) were conducted that also incorporated the elbow method on cross-entropy scores to determine the optimal number of clusters (K) (Fig. S1). Gene flow was assessed using a D-statistic approach based on sliding-windows (GitHub: simonhmartin/genomics_general) which was applied to the final SNP dataset. Non-overlapping sliding windows were set to a size of 100 kbp with a minimum of 100 SNPs. We assumed the topology (((WAus, Pac), Atl), Sei) to test for discrepancies in the frequencies of alternative topologies that would imply gene flow between the North-Atlantic population and either of the other two populations. For completeness, we also tested the two other possible topology combinations within the blue whales (Fig. S2). Additionally, we also used SambaR's "inferdemography()" function to calculate f2 and f3 statistics (following Patterson et al., 2012, Table S4) between all three blue whale populations.

Genetic differentiation and phylogeny

Raw genetic differences were calculated based on the total sets of variances (genomic, mitogenomic and mtMarker) including monomorphic sites using an inhouse script. The differences were then imported into SambaR to construct a BIO-neighbor-joining (BIONJ) tree (Gascuel, 1997) using the APE-5.3 package (Paradis & Schliep, 2019) and to generate a heatmap using ggplot2. Trees constructed for whole mitogenomic and sex-chromosome distances can be found in Fig. S3. The same steps were repeated for the haploid datasets based on the Y-chromosomal and mitogenomic variances.

Additionally, a dated phylogram (Fig. S4) was created based on the mitogenomic distances using SambaR's "popprtree()" function by specifying the sei whale as the outgroup and by providing a mutation rate of 4.0 × 10^{-8} bp^{-1} $year^{-1}$ (per site per year) (Brüniche-Olsen et al., 2021).

Genetic differentiation metrics, namely F_{ST} (Bhatia, Patterson, Sankararaman, & Price, 2013; Hudson, Slatkin, & Maddison, 1992) and d_A (Nei, 1987) were calculated using SambaR's "hudsonfst()" function using the raw genetic distance data calculated above. Doing so, we specified the Jukes-Cantor model as a substitution-model for correcting the nucleotide divergence metrics.

MSMC

Demographic models for all populations were constructed using the multiple sequential Markovian coalescent (MSMC2) framework (Schiffels & Wang, 2020). A mappability mask of the blue whale reference genome was created using SNPable (Heng Li, https://lh3lh3.users.sourceforge.net/snpable.shtml). To do this, we extracted 35-mers from the reference fasta file and mapped resulting sequences onto the reference genome using BWA aln with gap extension and gap opening penalties of 3. The mapping file was used to generate a mappability mask using SNPable's gen_mask function. Variant calling and the generation of individual masks for the "Super scaffold_1" was done with a combination of BCFTOOLS MPILEUP, BCFTOOLS CALL and the msmc-tools script bamCaller provided by the authors of MSMC2 (https://github.com/stschiff/msmc-tools) by using the same parameters as described above for individual variant calling and by piping the respective output into the bamCaller script. From these files, MSMC2 input files were generated using the msmc-tools script generate_multihetsep. Finally, MSMC2 was run with 100 Baumwelch iterations per individual and by using a reduced time segment patterning defined as "1*2+15*1+1*2" to avoid overfitting. To increase the resolution in the more recent time intervals (10 - 100 kya), we further tested phasing the input data with WHATSHAP v1.7 (Martin et al., 2016), more segments and more input scaffolds without changing the overall result.

Genetic diversity and inbreeding

Genome-wide heterozygosity was inferred by counting heterozygous sites within the individually inferred variance files that still include multivariate and monomorphic sites. An ANOVA test and a Tukey's post hoc test were performed to test for potentially significant differences.

All other genetic diversity parameters such as nucleotide diversity, Tajima's D and Watterson's Θ were calculated on a window-based approach using 100 kbp long, non-overlapping sliding windows. Nucleotide diversity was calculated per population using the general_genomics tool collection from Simon Martin (GitHub: simonhmartin/genomics_general) using thresholds that exclude windows if they contain less than 100 SNPs in more than 50% of all individuals. Neutrality tests around Tajima's D and Watterson's Θ were conducted using the ANGSD v0.931 tool collection following the user recommendations (Korneliussen, Albrechtsen, & Nielsen, 2014). Folded side frequency spectra (fSFS), necessary to calculate the number of segregation sites, were directly inferred from mapping files using a combination of the "-doSaf 1", the "realSFS -fold 1", the "saf2theta" and "thetaStat" functions. A more detailed description of the used methods can be found in (Korneliussen, Moltke, Albrechtsen, & Nielsen, 2013).

Runs of homozygosity were identified in the combined set of autosomal variances with DARWINDOW (https://github.com/mennodejong1986/Darwindow), by using a sliding window approach to determine levels of heterozygosity per 20kbp window. We further excluded scaffolds smaller than 5Mbp and plotted the levels of heterozygosity over all inferred scaffolds (example for Super scaffold_9 depicted in Fig S5). The minimum number of consecutive windows and the heterozygosity threshold per window were adjusted until we reached the best fit between visible ROHs and ROHs defined by DARWINDOW. This resulted in a minimum number of 50 windows and hence a minimum ROH length of 1Mbp. The heterozygosity threshold reached a best fit at 0.06 for most individuals, although for four particularly heterozygous individuals, a heterozygosity threshold of 0.12 resulted in the best fit. ROHs were then sorted into eleven different length (Mbp) bins (1-1.5, 1.5-2, 2-2.5, 2.5-3, 3-3.5, 3.5-4, 4-5,5-6,6-8,8-10,10-open). Per bin, inbreeding coefficients (F_{ROH}, after McQuillan et al. 2012) were calculated as the proportion of the reference genome covered by the sum of all corresponding ROHs. A one-way ANOVA test and a Tukey's post hoc test were used to detect significant differences between F_{ROH} bins of different populations, although tests were only possible until bin number 7 (4-5Mbp) as later bins did not include enough data to make meaningful comparisons.

Results

Whole-genome short-read datasets were compiled for 20 blue whale individuals originating from the northern and equatorial Pacific and Western-Australian waters (Fig. 1), as well as for one additional Sei whale individual from Icelandic waters. Per individual, we generated on average 477 million short-read pairs totaling to a mean of 72 Giga base pairs (Gbp) of sequence data per sample, equaling a 30-fold coverage. This dataset was complemented by publicly available sequences of 11 North-Atlantic blue whales, another North-Pacific individual and two Sei whale genomes resulting in a total set of 35 individuals represented by genome data (Table S1, S2). From this, we collected a sampling-wide set of 1.5 million high-quality SNPs that were used for downstream analyses. Additionally, we extracted mitogenomic sequences of all included individuals for additional analyses.

Population structure

Estimates of ancestry coefficients and population structure were inferred using the R LEA package assuming two to six ancestral populations (K) (Fig. 2A, Fig. S1). Minimal cross-entropy scores increased with K and were unable to unambiguously identify a best fitting model using the elbow method. When assuming two populations (K = 2), a clear separation between the Western-Australian pygmy blue whales and the two other sampled northern-hemisphere oceans is evident. At K = 3, the two northern populations are further differentiated into an Atlantic and a Pacific population, consistent with the next-lowest cross-entropy score. A higher number of ancestral populations did not result in any biological meaningful clusters. A principal coordinate analysis (PCoA, Fig. 2B) confirmed the existence of three distinct blue whale populations with the two first axes of differentiation. The PC1 explained 33.2% of the variation and separated the pygmy blue whales from northern blue whales while the PC2 explained 21.1% of the variation and separated Atlantic from Pacific individuals.

Genetic exchange

Signs of admixture were limited in the admixture analysis and mostly non-consistent over different numbers of inferred populations (Fig. 2A). In the K = 2 inference, all Pacific individuals showed minimal signs of admixture with the pygmy blue whale, which only remain, albeit to a lesser extent, in three individuals over higher values of K. One Atlantic individual (BM1401) depicts a more prominent signal of admixture with Pacific individuals. This signal is consistent over all numbers of K.

Gene flow analyses using a sliding-window based D-statistic "ABBA-BABA" approach resulted in positive D values and Z-scores over 3 in all possible inferred topology-combinations, using the Sei whale as an outgroup (Fig. 2C, Fig. S2). When using the Atlantic population as the introgressor, a D value of 0.026 indicates an excess of genetic signals with the Pacific population compared to the Australian population (Fig. 2C). Placing the Pacific population outside of the Atlantic and Australian populations resulted in a lower but still positive value of D = 0.016, suggesting a connectivity between Pacific and Western-Australian individuals (Fig. S2A). A grouping of the two northern populations using the Australian populations as introgressor displayed the highest value of D = 0.043, suggesting an excess of signals between the Pacific and Western-Australian group as well (Fig. S2B). Additionally, a calculation of f3-statistics resulted in positive f3 values with high Z scores and significant p-values regardless of the tested combination, indicating no detectable signals in one of the populations that would support an intermediate origin from between the other two populations (Table S3B).

Genetic differentiation

Genomic differentiation estimates using F_{ST} statistics (Bhatia et al., 2013; Hudson et al., 1992), net nucleotide divergence d_A (Nei, 1987) (Table 1), and f2 statistics (Patterson et al., 2012) (Table S3A), suggested a higher fixation rate but low divergence between all blue whale populations. F_{ST} values between blue whale populations ranged from 0.06 - 0.15, with Atlantic and Pacific having the lowest (0.062) and Atlantic and West-Australia having the highest (0.147) genetic differentiation. Net nucleotide divergence (d_A), a measure of genetic divergence corrected for within-group genetic diversity, was again smallest in the comparison of the two northern populations (0.014%), and higher in comparisons with the Western-Australian blue whales (0.039% and 0.032%). The f2-statistics, denoting the mean squared difference between allele frequencies of two populations, yielded values between 0.02 and 0.05 (Table S3A?). While the lowest differentiation was found between the Atlantic and Pacific population (f2 = 0.02), we noted higher values between each of both northern populations and the Western-Australian pygmy blue whale population (Atl-WAus = 0.051; Pac-WAus = 0.043).

Differentiation statistics of the mitochondrial genomes generally coincided with their genome-wide equivalents but found a lower divergence between Pacific and Australian blue whales, comparable or even lower than between Atlantic and Pacific. F_{ST} values ranged between 0.11 - 0.227, with the Atlantic and Pacific pair having the lowest fixation rate, followed by the pair of Pacific and Australian blue whales (0.145) and capped off by the Atlantic and Australian pair. Net nucleotide divergence, d_A was lowest between the Pacific and Australian individuals (0.0151%). Atlantic and Pacific individuals featured a d_A of 0.0196% and Atlantic and Australian individuals diverged by 0.0362%.

Since past comparisons were often based on the mitochondrial marker sequence (See (Rosel et al., 1994), we also extracted this specific region of from our mitogenomic set of variances and rerun the analysis with similar results. F_{ST} values were between 0.071 – 0.313, again with Atlantic and Pacific having the lowest fixation rate and Atlantic and West-Australia having the highest fixation rate. Similar to the whole genome results, Atlantic and Pacific featured the least genome wide divergence with d_A = 0.086%, while Atlantic and Australian blue whales were most diverged with d_A = 0.431%. Like for the whole genome results and contrary to the whole mitogenome values, the pair of Pacific and Western Australia featured intermediate divergence with d_A = 0.24%.

Genetic distance and phylogeny

Pairwise genetic distances were calculated based on the total set of variances including monomorphic sites. The phylogenetic tree based on BIONJ clustering (Gascuel, 1997) resolved all three blue whale populations as in the population structure analyses and displayed a genetic distance of about 0.8% between the Sei whale and the inferred blue whale populations (Fig 3A). The unrooted topology groups North-Pacific and North-Atlantic individuals, as well as Western-Australian individuals and the Sei whale outgroup, although both to a minimal extent. Distances between populations were small compared to the species level distance, while West-Australian pygmy blue whale individuals showed the highest level of population-specific signals compared to the two northern populations.

A heatmap created using the pairwise genetic distances directly depicts a similar situation (Fig 3B). More specifically, the distance between the North-Atlantic blue whale and the West-Australian pygmy whale population was the highest (~0.024%), followed by North-Pacific and West-Australia (~0.023%) and formally capped off by the lowest distance between North-Atlantic and North-Pacific (~0.021%). Interpopulation distances were lower in general and lowest between Western-Australian individuals. Furthermore, we found two outlier individuals with higher amounts of individual-specific genetic signals, namely individual 5810 (Pac) and 23982 (WAus), which also appear distinctly in other analyses.

Haploid datasets, retrieved from mitochondrial and Y-chromosomal specific loci resulted in trees with less distinct clusters (Fig. S3). The unrooted mitochondrial tree finds three main clusters that are roughly consistent with the population assignment, although they, in some instances, group individuals from different localities. Furthermore, multiple more distant clusters of Atlantic and Pacific individuals were found that also appeared to be more ancient in a dated phylogeny (Fig. S4). The split between the three main clusters was estimated to have occurred between 10,000 and 15,000 years ago. The unrooted Y-chromosomal tree did not show meaningful differentiation between Atlantic and Pacific individuals while the Western-Australian individuals formed a distinct clade that also includes the sei whale outgroup.

Demography

Demographic models for each individual were generated using the MSMC2 framework (Schiffels & Wang, 2020) and were depicted per population in Fig. 4. We were able to infer coalescent events within a timeline of 10 million years (Mya) to 100 thousand years ago (kya), spanning over a time window that includes the two major oceanic transition events, the Mid-Pleistocene transition (MPT, 1.25 Mya to 700 kya, Clark et al., 2006) and the Late-Pleistocene transition (LPT, 3.2 Mya to 2.6 Mya, Hill, Bolton, & Haywood, 2017). All models, including the Sei whale outgroup, depict a peak in the effective population size at earlier modeled times, with blue whale models having a peak at around 4-5 Mya and the three sei whale models indicating a peak at around 2-3 Mya. All blue whale inferences suggest a stable population size within a time-window of 1.5 Mya to 300 kya years ago, before starting to decline again. The Western-Australian population also indicates a local peak at around 100 kya before reaching a stable, lower population size compared to the two northern populations. Apart from this, differences in the demographic trajectory between the three blue whale subspecies were marginal which does not allow to infer a split between them.

Genomic diversity and inbreeding

Genomic diversity was assessed by the means of genome-wide heterozygosity, nucleotide diversity (π), Tajima's D, Watterson's Θ and inbreeding coefficients based on runs of homozygosity (F_{ROH}) (Fig. 5, Table 2). Heterozygosity was comparable between the two northern populations with a mean value of 0.20% and 0.21% in the Atlantic and Pacific population, respectively (Fig. 5A). Australian pygmy blue whales showed the highest variance and the highest mean value of 0.22% but were not significantly higher compared to the other populations (ANOVA: f = 0.94, p = 0.4; Tukey: p = 0.4). Three individuals were excluded due to their exceptionally high level of heterozygosity that might have been caused by hybridization with a non-sampled population, or another species rather than representing the heterozygosity of the respective population: Pacific: 5810; Australian: 23982 and 42284. Nucleotide diversity, defined as the mean pairwise difference within individuals of a population, was found to closely resemble the overall heterozygosity levels, with the Pacific and West-Australian population having slightly lowered values compared to mean heterozygosity (Pac = 0.20%, WAus = 0.21%).

Neutrality tests, meant to distinguish whether a population is evolving neutrally or not, were based on the folded side frequency spectrum (fSFS) (See (Korneliussen et al., 2013) to estimate the number of segregating sites. This test revealed a similar positive Tajima's D for the Atlantic and Western-Australian population (Atl = 0.88, WAus = 0.86) and a lower but still positive value for the Pacific population (Pac = 0.59), indicating an excess of variation in the observed diversity compared to the expected diversity in all inferred groups. Watterson's Θ, a measure of the expected nucleotide diversity within a group or population, was similar in all three populations and ranged between 121.1 (Atl), 121.3 (Pac) and 122.4 (WAus).

Inbreeding coefficients showed a decreasing pattern with increasing ROH size and no prominent nor significant differences between the blue whale populations (Fig 5B-D). Pacific individuals featured the lowest amount of short ROHs (<2 Mbp) and the overall longest ROHs with a record of a 17 Mbp long ROH in the individual 17960, suggesting influences of more recent inbreeding. By contrast, the Australian pygmy blue whales had the highest amounts of short ROHs and did not feature ROHs longer than 5.5 Mbp.

Discussion

In this study we compared genomes of blue whales from the northern Atlantic and northern Pacific blue whale to the well-established pygmy blue whale subspecies and find an overall high degree of isolation, but a surprisingly limited and contradicting low genomic and mitogenomic divergence. Given that the pygmy blue whale is an established, morphologically different, and well recognized subspecies, we propose that the two northern populations represent equally independent branches of the overall distribution of the blue whale.

Isolation

We report a clear separation among the three populations that coincides with their geographic distance. A northern connection between Atlantic and Pacific blue whale populations has not been reported thus far and seems unlikely given their dependence on a high ocean productivity in areas connected to deep sea ocean dynamics (Branch et al., 2007). This indicates a long geographic distance between the Atlantic and Pacific populations which is comparable or even larger than to towards the Australian pygmy blue whales. Evidence of genetic exchange was limited among all three groups with, compared to the northern Atlantic population, a slightly higher connectivity between the northern Pacific population and the Australian pygmy blue whale population. Considering the lack of ocean productivity in the waters of the central Pacific, we assume that the most likely route for genetic exchange between all three populations is via the Antarctic Ocean inhabited by the southern blue whale subspecies (*B. m. intermedia*) (Branch, T. A. et al., 2007). This hypothesis is at least partially supported by reports of a hybridization zone between the Australian pygmy blue whales and the Antarctic subspecies (Attard et al., 2012) as well as the occurrence of Antarctic blue whales in the Southeast Pacific (Leduc et al., 2017). A sophisticated test of this hypothesis requires, however, genome data from individuals from the Antarctic regions and will be addressed in future studies.

Differentiation

Genomic pairwise nuclear fixation indices (F_{ST}: 0.062 - 0.14) between the three blue whale populations, including the here sampled Australian pygmy blue whales, fall into the range of other recognized *Cetacea* subspecies, like 0.004 - 0.012 for the spinner dolphins (*Stenella longirostris*) (Leslie & Morin, 2018) and 0.18 - 0.24 for the Harbor porpoise (*Phocoena phocoena*) (Lah et al., 2016), but also into the range of other proposed isolated populations, like 0.09 for sympatric killer whale populations (*Orcinus orca*) (Foote et al., 2016) and 0.018 – 0.197 for the Cuvier's beaked whale (Ziphius cavirostris) (Onoufriou et al., 2022). Same is found for the mitogenomic and mitochondrial control region F_{ST} values compared to other *Odontoceti* subspecies (Leslie and Morin, 2018) and isolated populations (Onoufriou et al., 2022) but were generally higher compared to the arguably more similar fin whale subspecies (*Balaenoptera physalus subspp.*, Archer et al., 2019). The genomic and mitogenomic net nucleotide divergence (genomic d_A: 0.00014 – 0.00039, mitogenomic d_A: 0.00015 – 0.00036), however, was drastically lower when compared to other comparable data of whales, like a mitogenomic d_A of 0.0097 for Cuvier's beaked whales from different ocean basins (Onoufriou et al., 2022) and a mitogenomic d_A of 0.0036 – 0.0072 for fin whales from different ocean basins (Archer et al., 2019). Using only the control region locus of the mitogenomic sequences resulted in generally low divergence values too (d_A: 0.0009 – 0.0043). Only the arguably most distant pair, namely North Atlantic and Western-Australia, barely met a suggested threshold of $d_A > 0.004$, which formally defines a lower boundary of declaring subspecies (Rosel et al., 2017). All other pairs, including the pair of Western-Australia and North Pacific, did not met this proposed threshold.

A similar picture is found in analyses of genome-wide distances as there is an overall low genetic distance compared to the sei whale with a slightly increased population specific distance in the Australian group. Congruent with the gene flow and admixture analyses, all divergence and distance statistics indicate a higher degree of differentiation between the North Atlantic blue whale to the Australian pygmy blue whale population compared to the North Pacific and Australian population. Distance based trees constructed with haploid data like mitochondrial or Y-chromosomal genotypes, however, revealed a more admixed past, given that these loci do not recombine and showed less clear or missing clustering between the populations. Divergence times estimated by mitochondrial data indicate a main split between 10,000 - 15,000 years ago, although some northern-hemisphere individuals fell out of these main groups, indicating a potential earlier colonization of these oceanic regions. The divergence time estimation is roughly consistent with the estimated divergence time between Australian and Antarctic blue whales some 20,000 years ago (Attard et al., 2015). Y-chromosomal data furthermore depicts a lack of resolution between both northern populations and may therefore indicate male-biased genetic exchange. Nevertheless, our genome-wide gene flow analyses oppose a strong influence of this phenomenon.

Subspecies inference

In a special issue of Marine Mammal Science (Volume33, IssueS1), Taylor and colleagues proposed guidelines for taxonomic decisions for cetacean species that should ensure consistency when using molecular data (Taylor, Archer et al., 2017; Taylor, Perrin et al., 2017). Within this work, they formulated a subspecies definition as an "... independently evolving unit that appears to be on its way towards speciation" (Taylor, Perrin et al., 2017). Below the subspecies level, they proposed two smaller units, namely a demographically independent population (DIP) and an evolutionarily significant unit (ESU) characterized by different degrees of isolation without being diagnosable different. To guide decisions based on mitochondrial control region data, they suggested a threshold of net nucleotide divergence for subspecies as $d_A > 0.004$, a level of divergence which was not met in our data, including the already established and morphologically different pygmy blue whale (Ichihara, 1966). DIP's can be excluded based on the high degree of isolation between all oceans, leaving ESU's as the best fit for all blue whale populations given the statistical context provided by Rosel et al., 2017. However, in the case of a potential recent split in combination with other lines of evidence, these guidelines may still suggest a subspecies separation. A recent split after the last glacial maximum (LGM) seems plausible considering the here estimated split based on mitochondrial data that coincides with other demographic assessments of the Australian pygmy blue whale and the Antarctic blue whale, which estimated their split at around 20,000 years ago (Attard et al., 2015). The larger ice caps that existed during the LGM might have made many current summer foraging areas inaccessible as well as the marine ecosystems less productive in general (Cabrera et al., 2022). This might have facilitated a greater need for migration and a greater admixture between different populations as already discussed in the regards of the evolution of baleen whale migrations in general (Slater, Goldbogen, & Pyenson, 2017). Given that the size of the blue whale allows for extremely long possible travel distances, the potential for admixture might have been especially high in blue whales, resulting in the low genetic divergence that only started to accumulate after previously high latitude habitats became available again in the Pleistocene-Holocene transition 12,000-7,000 years ago (Cabrera et al., 2022).

Morphological differences between both populations are not known but might be existent given that no direct comparison between both groups was attempted thus far. The reasons for this might simply be logistical challenges that arise from the enormous body sizes and the general inaccessibility of pelagic animals, as outline by Taylor, Perrin et al., 2017. Other phenotypic lines of evidence for a differentiation into subspecies are acoustic patterns that enable a clear distinction between songs of whales from both northern oceans (McDonald et al., 2006; Mellinger & Clark, 2003). Songs of Northern Atlantic blue whales include characteristic tonal, single-phrased, very-low-frequency calls in the 15–20 Hz range with two units of which the latter may sweep to even lower frequencies (McDonald et al., 2006; Mellinger & Clark, 2003). Northern Pacific blue whales have one of two characteristic song patterns specific for the two respective major populations in the northern Pacific (McDonald et al., 2006; Stafford, Nieukirk, & Fox, 2001). The well-studied eastern North Pacific song pattern consists of at least two phrases with the first one being pulsed with multiple, time-offset non-harmonic components and the latter one being tonal with a series of harmonically related higher frequencies (McDonald et al., 2006). It is noteworthy that a decline of acoustic frequencies in both phrases has been reported since the 1960s (Rice et al., 2022). Central North Pacific individuals feature songs with 2-4 tonal units at the 20 Hz band with varying usage (Stafford et al., 2001).

Overall, the genomic differences between the two northern hemisphere blue whale populations are comparable to that of the relatively well-established pygmy blue whale and call for an equal recognition in the scientific community. Hence the decision is left to either revoke the subspecies status of the pygmy blue whale or to establish new subspecies. Given that these levels of genetic differences led to morphologically distinct features in the pygmy blue whale and already reported phenotypic acoustic traits that distinguish both populations, we would support the latter of both possibilities. Eventually, we also expect a subspecies-recognition to have beneficial effects on their received conservation efforts as discussed in Taylor, Perrin et al., 2017.

Subspecies definition

In the following, we propose an idea for new subspecies of the two major northern hemisphere blue whale populations. In accordance with the International Code of Zoological Nomenclature (ICZN) priority principle, we would leave the name for the North Atlantic subspecies as *Balaenoptera musculus musculus* due to the first scientific name of the blue whale formulated by Carl von Linné in 1758 (Linné, 1758). Diagnostic features for the Atlantic blue whale (*B. m. musculus*) would be its song pattern (Clark and Mellinger et al. 2003, McDonald et al. 2006) and their geographic range in the North-East Atlantic around the waters of West Greenland, Iceland, Norway, Ireland, the Shetland Islands, the Hebrides, and the Faroes Islands and in the North-West Atlantic in Canadian waters around the Gulf of St. Lawrence, the Scotian Shelf, the Bay of Fundy as well as in US waters around Cape Cod, MA) (Lesage, V., Gosselin, J.-F., Lawson, J.W., McQuinn, I., Moors-Murphy, H., Plourde, S., Sears, 2018; Pike, Víkingsson, Gunnlaugsson, & Øien, 2013). Since a North Pacific blue whale specimen has already been described as *Balaenoptera musculus sulfureus* by Cope in 1869, we propose to reestablish this name (Cope 1869). Diagnostic features for the Pacific blue whale (*B. m. sulfureus*) would be its geographic distribution around the North Pacific coasts (Gilpatrick Jr. & Perryman, 2008), and one of two characteristic song patterns specific for individuals of either the central North Pacific or the eastern North Pacific population (McDonald et al., 2006; Stafford et al., 2001).

In the case of the North Pacific blue whale (*B. m. sulfureus*), we like to emphasize that LeDuc and colleagues found reasonable evidence that North (-east) Pacific and Southeast Pacific individuals were more similar to each other compared to either the pygmy blue whale (*B. m. brevicauda*) or the Antarctic blue whale (*B. m. intermedia*) (Leduc et al., 2017). Hence, we expect a revision of the distribution of the North Pacific blue whale (*B. m. sulfureus*), as soon as more sophisticated comparisons are available to account for these findings.

Conservation genomics

The molecular viability assessment of all included blue whale populations was conducted to evaluate their genetic diversity and levels of inbreeding to provide a genomic basis for conservation efforts and to mitigate the impact of 20th century industrial whaling.

Consistent with previous studies, we reported high levels of genome-wide heterozygosity when compared to other baleen whales (Árnason, Lammers, Kumar, Nilsson, & Janke, 2018; Jossey et al., 2021; Wolf et al., 2022) and other mammals (Brüniche-Olsen, Kellner, Anderson, & DeWoody, 2018; Palkopoulou et al., 2015). Although it may seem contradictory at first due to the expected enormous decimation during the industrial whaling period (Tønnessen & Johnsen, 1982), we would like to emphasize that industrial whaling occurred two to three generations ago, making a manifestation in the genotype unlikely. Furthermore, levels of heterozygosity have been described as a poor indicator for the IUCN red list status (Brüniche-Olsen et al., 2018; Teixeira & Huber, 2021) and that other factors like, population size, physiological parameters and the long-term demographic history are more determinant for the level of heterozygosity than IUCN status (Brüniche-Olsen, Kellner, & DeWoody, 2019; Charlesworth & Jensen, 2022). Like proposed previously, we therefore expect other measures to be more informative about potential negative consequences for the reduced stock sizes.

A test for evolutionary neutrality in our blue whale populations resulted in positive Tajima's D statistics that indicates a higher number of sequences that deviate from neutral evolution. More specifically, a positive value indicates a higher divergence of pairwise genetic difference within a population that is higher than one could expect. Hence a positive D value indicates a lack of alleles at lower frequencies. Two main scenarios can result in a positively deviating Tajima's D (Schmidt & Pool, 2002). First, balancing selection that actively maintains multiple alleles or second, a recent population contraction (Schmidt & Pool, 2002). Because the impact of whaling on stocks sizes was sometimes so severe that local industries collapsed or had to switch to smaller baleen whale species due to economically non-sufficient catch rates (Tønnessen & Johnsen, 1982), we assume that the latter reason is at least a major driver of this genome-wide shift away from evolutionary neutrality.

Signs of inbreeding measured by the means of runs of homozygosity over one million base-pairs were found in all individuals except some highly heterozygous and potentially outbred outliers. Although we did not find major differences between populations, these results highly diverge from comparable analyses done on an Icelandic fin whale population (Wolf et al., 2022). Comparing differences between runs of homozygosity results are often problematic due to many variable factors induced by different software, definition settings and reference genome qualities (Prasad, Lorenzen, & Westbury, 2022). However, in this case, both analyses were done with comparable data, identical approaches, and both identified ROH by manual verification. Thus, the nearly tenfold higher inbreeding parameters in the three blue whale populations compared to the Icelandic fin whale population, may in fact be indicative of increased inbreeding in all blue whale populations. These results are congruent with reports of a more impactful population decline of blue whales by whaling compared to fin whales (Tønnessen & Johnsen, 1982).

Although the blue whale appears to recover in general, our found genome-wide consequences implicate that careful monitoring and conservation efforts, including more comprehensive genome analyses, are nonetheless necessary to ensure the persistence of its recovery. There is a possibility, that these genomic changes may have long-term effects on the general fitness of the species due to the reciprocal relationship between inbreeding, genetic diversity, mutational load and population sizes, also called “inbreeding depression”, “mutational meltdown” or “extinction vortex” (Teixeira & Huber, 2021). Especially constant inbreeding could lead to a fixation of previously heterozygous recessive mutations in emerging runs of homozygosity (Teixeira & Huber, 2021; van der Valk, Manuel, Marques-Bonet, & Guschanski, 2019; Wolf et al., 2022) that would affect the long-term fitness of the blue whale.

Conclusion

Our population genomic analyses made it possible to assess the degree of isolation and the genome-wide divergence between the two major populations of the northern hemisphere blue whale subspecies (*B. m. musculus*) and compare the results to the established pygmy blue whale subspecies (*B. m. brevicauda*). Because of the high degree of isolation and low but comparable genomic differences between all three groups, we propose to separate Northern Pacific and Northern Atlantic blue whale into two subspecies (*B. m. musculus, B. m. sulfureus*). Our molecular viability assessment also revealed a significant impact of inbreeding on blue whale genome together with the loss of rare genetic alleles that may, together with the new subspecies status, inspire enhanced conservation efforts to protect this iconic species.

Acknowledgments

We are grateful to Kelly Robertson and to the NOAA Southwest Fisheries Science Center MMaSTR Collection, La Jolla, California, USA, for their sampling efforts, for providing DNA samples, and for their general research assistance. We thank Aimeé Lang for aiding with the writing process and helpful taxonomic insights. Furthermore, we thank Dr. Carola Greve and the TBG Lab Center Team for their assistance with the sequencing process and Marcel Nebenfuehr for his support in general lab work regarding the sei whale sample. Eventually, we thank Jón Baldur for providing the blue whale illustration.

Funding

The present study was funded by the Centre for Translational Biodiversity Genomics (LOEWE-TBG) through the program "LOEWE – Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts.

Competing interests

The authors declare that they have no competing interests.

References

Archer, F. I., Brownell, R. L., Hancock-Hanser, B. L., Morin, P. A., Robertson, K. M., Sherman, K. K., . . . Moratelli, R. (2019). Revision of fin whale Balaenoptera physalus (Linnaeus, 1758) subspecies using genetics. *Journal of Mammalogy*, *100*(5), 1653–1670. https://doi.org/10.1093/jmammal/gyz121

Árnason, Ú., Lammers, F., Kumar, V., Nilsson, M. A., & Janke, A. (2018). Whole-genome sequencing of the blue whale and other rorquals finds signatures for introgressive gene flow. *Science Advances*, *4*(4), eaap9873. https://doi.org/10.1126/sciadv.aap9873

Attard, C. R. M., Beheregaray, L. B., Jenner, K. C. S., Gill, P. C., Jenner, M.-N., Morrice, M. G., . . . Möller, L. M. (2012). Hybridization of Southern Hemisphere blue whale subspecies and a sympatric area off Antarctica: Impacts of whaling or climate change? *Molecular Ecology*, *21*(23), 5715–5727. https://doi.org/10.1111/mec.12025

Attard, C. R. M., Beheregaray, L. B., Jenner, K. C. S., Gill, P. C., Jenner, M.-N. M., Morrice, M. G., . . . Möller, L. M. (2015). Low genetic diversity in pygmy blue whales is due to climate-induced diversification rather than anthropogenic impacts. *Biology Letters*, *11*(5), 20141037. https://doi.org/10.1098/rsbl.2014.1037

Bhatia, G., Patterson, N., Sankararaman, S., & Price, A. L. (2013). Estimating and interpreting FST: The impact of rare variants. *Genome Research*, *23*(9), 1514–1521. https://doi.org/10.1101/gr.154831.113

Branch, T. A., Abubaker, E. M. N., Mkango, S., & Butterworth, D. S. (2007). SEPARATING SOUTHERN BLUE WHALE SUBSPECIES BASED ON LENGTH FREQUENCIES OF SEXUALLY MATURE FEMALES. *Marine Mammal Science*, *23*(4), 803–833. https://doi.org/10.1111/j.1748-7692.2007.00137.x

Branch, T. A., Stafford, K. M., Palacios, D. M., Allison, C., Bannister, J. L., Burton, C. L.K., . . . Warneke, R. M. (2007). Past and present distribution, densities and movements of blue whales Balaenoptera musculus in the Southern Hemisphere and northern Indian Ocean. *Mammal Review*, *37*(2), 116–175. https://doi.org/10.1111/j.1365-2907.2007.00106.x

Brüniche-Olsen, A., Bickham, J. W., Godard-Codding, C. A., Brykov, V. A., Kellner, K. F., Urban, J., . . . Smyser, T. (2021). Influence of Holocene habitat availability on Pacific gray whale (Eschrichtius robustus) population dynamics as inferred from whole mitochondrial genome sequences and environmental niche modeling. *Journal of Mammalogy*, *102*(4), 986–999. https://doi.org/10.1093/jmammal/gyab032

Brüniche-Olsen, A., Kellner, K. F., Anderson, C. J., & DeWoody, J. A. (2018). Runs of homozygosity have utility in mammalian conservation and evolutionary studies. *Conservation Genetics*, *19*(6), 1295–1307. https://doi.org/10.1007/s10592-018-1099-y

Brüniche-Olsen, A., Kellner, K. F., & DeWoody, J. A. (2019). Island area, body size and demographic history shape genomic diversity in Darwin's finches and related tanagers. *Molecular Ecology*, *28*(22), 4914–4925. https://doi.org/10.1111/mec.15266

Bukhman, Y. V., Morin, P. A., Meyer, S., Chu, L.-F., Jacobsen, J. K., Antosiewicz-Bourget, J., . . . Stewart, R. (2022). *A high-quality blue whale genome, segmental duplications, and historical demography.*

Burtenshaw, J. C., Oleson, E. M., Hildebrand, J. A., McDonald, M. A., Andrew, R. K., Howe, B. M., & Mercer, J. A. (2004). Acoustic and satellite remote sensing of blue whale seasonality and habitat in the Northeast Pacific. *Deep Sea Research Part II: Topical Studies in Oceanography*, *51*(10-11), 967–986. https://doi.org/10.1016/j.dsr2.2004.06.020

Cabrera, A. A., Schall, E., Bérubé, M., Anderwald, P., Bachmann, L., Berrow, S., . . . Palsbøll, P. J. (2022). Strong and lasting impacts of past global warming on baleen whales and their prey. *Global Change Biology*, *28*(8), 2657–2677. https://doi.org/10.1111/gcb.16085

Charlesworth, B., & Jensen, J. D. (2022). How Can We Resolve Lewontin's Paradox? *Genome Biology and Evolution*, *14*(7). https://doi.org/10.1093/gbe/evac096

Chen, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics (Oxford, England)*, *34*(17), i884-i890. https://doi.org/10.1093/bioinformatics/bty560

Clark, P. U., Archer, D., Pollard, D., Blum, J. D., Rial, J. A., Brovkin, V., . . . Roy, M. (2006). The middle Pleistocene transition: Characteristics, mechanisms, and implications for long-term changes in atmospheric pCO2. *Quaternary Science Reviews*, *25*(23-24), 3150–3184. https://doi.org/10.1016/j.quascirev.2006.07.008

Cooke, J. G. (2018). Blue Whale: Balaenoptera musculus. *The IUCN Red List of Threatemed Species*.

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., . . . Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., . . . Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). https://doi.org/10.1093/gigascience/giab008

Double, M. C., Andrews-Goff, V., Jenner, K. C. S., Jenner, M.-N., Laverick, S. M., Branch, T. A., & Gales, N. J. (2014). Migratory movements of pygmy blue whales (Balaenoptera musculus brevicauda) between Australia and Indonesia as revealed by satellite telemetry. *PloS One*, *9*(4), e93578. https://doi.org/10.1371/journal.pone.0093578

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

Frichot, E., & François, O. (2015). LEA: An R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, *6*(8), 925–929. https://doi.org/10.1111/2041-210X.12382

Gascuel, O. (1997). BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, *14*(7), 685–695. https://doi.org/10.1093/oxfordjournals.molbev.a025808

Gilpatrick Jr., J. W., & Perryman, W. L. (2008). Geographic variation in external morphology of North Pacific and Southern Hemisphere blue whales (Balaenoptera musculus). *J. Cetacean Res. Manage.*, *10*(1), 9–21. https://doi.org/10.47536/jcrm.v10i1.654

Hill, D. J., Bolton, K. P., & Haywood, A. M. (2017). Modelled ocean changes at the Plio-Pleistocene transition driven by Antarctic ice advance. *Nature Communications*, *8*, 14376. https://doi.org/10.1038/ncomms14376

Hucke-Gaete, R., Bedriñana-Romano, L., Viddi, F. A., Ruiz, J. E., Torres-Florez, J. P., & Zerbini, A. N. (2018). From Chilean Patagonia to Galapagos, Ecuador: Novel insights on blue whale migratory pathways along the Eastern South Pacific. *PeerJ*, *6*, e4695. https://doi.org/10.7717/peerj.4695

Hudson, R. R., Slatkin, M., & Maddison, W. P. (1992). Estimation of levels of gene flow from DNA sequence data. *Genetics*, *132*(2), 583–589. https://doi.org/10.1093/genetics/132.2.583

Ichihara, T. (1966). The pygmy blue whale, Balaenoptera musculus brevicauda, a new subspecies from the Antarctic. In K. S. Norris (Ed.), *Library reprint series. Whales, Dolphins and Porpoises: Procs. of symp*. S.l.: Univ of California Press.

Jong, M. J. de, Jong, J. F. de, Hoelzel, A. R., & Janke, A. (2021). SambaR: An R package for fast, easy and reproducible population-genetic analyses of biallelic SNP data sets. *Molecular Ecology Resources*, *21*(4), 1369–1379. https://doi.org/10.1111/1755-0998.13339

Jossey, S., Haddrath, O., Loureiro, L., Lim, B., Miller, J., Lok, S., . . . Engstrom, M. (2021). *Blue whale (Balaenoptera musculus musculus) genome: Population structure and history in the North Atlantic*.

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of Next Generation Sequencing Data. *BMC bioinformatics*, *15*(1), 443. https://doi.org/10.1186/s12859-014-0356-4

Korneliussen, T. S., Moltke, I., Albrechtsen, A., & Nielsen, R. (2013). Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC Bioinformatics, 14*, 289. https://doi.org/10.1186/1471-2105-14-289

Lah, L., Trense, D., Benke, H., Berggren, P., Gunnlaugsson, Þ., Lockyer, C., . . . Tiedemann, R. (2016). Spatially Explicit Analysis of Genome-Wide SNPs Detects Subtle Population Structure in a Mobile Marine Mammal, the Harbor Porpoise. *PloS One, 11*(10), e0162792. https://doi.org/10.1371/journal.pone.0162792

Leduc, R. G., Archer, F. I., Lang, A. R., Martien, K. K., Hancock-Hanser, B., Torres-Florez, J. P., . . . Taylor, B. L. (2017). Genetic variation in blue whales in the eastern pacific: Implication for taxonomy and use of common wintering grounds. *Molecular Ecology, 26*(3), 740–751. https://doi.org/10.1111/mec.13940

Leduc, R. G., Dizon, A. E., Goto, M., Pastene, L. A., Kato, H., Nishiwaki, S., . . . Brownell, R. L. (2007). Patterns of genetic variation in Southern Hemisphere blue whales and the use of assignment test to detect mixing on the feeding grounds. *J. Cetacean Res. Manage., 9*(1), 73–80. https://doi.org/10.47536/jcrm.v9i1.694

Lesage, V., Gosselin, J.-F., Lawson, J.W., McQuinn, I., Moors-Murphy, H., Plourde, S., Sears (2018). *Habitats important to blue whales (Balaenoptera musculus) in the western North Atlantic.* Doc. 2016/080. iv + 50 p: DFO Can. Sci. Advis. Sec. Res.

Leslie, M. S., & Morin, P. A. (2018). Structure and phylogeography of two tropical predators, spinner (Stenella longirostris) and pantropical spotted (S. attenuata) dolphins, from SNP data. *Royal Society Open Science, 5*(4), 171615. https://doi.org/10.1098/rsos.171615

Linné, C. v. (1758). *Caroli Linnaei…Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis*. Holmiae: Impensis Direct. Laurentii Salvii.

Ljunglad, D. K., Clark, C. W., & Shimada, H. (1998). A comparison of sounds attributed to pygmy blue whales (Balaenoptera musculus brevicauda) recorded south of the Madagascar plateau and thos attributed to 'true' blue whales (Balaenoptera musculus) recorded off Antartica. *Rep. Int. Whal. Commn., 48*, 439–444.

Martin, M., Patterson, M., Garg, S., O Fischer, S., Pisanti, N., Klau, G. W., . . . Marschall, T. (2016). WhatsHap: fast and accurate read-based phasing. *bioRxiv, 085050*.

McClain, C. R., Balk, M. A., Benfield, M. C., Branch, T. A., Chen, C., Cosgrove, J., . . . Thaler, A. D. (2015). Sizing ocean giants: Patterns of intraspecific size variation in marine megafauna. *PeerJ, 3,* e715. https://doi.org/10.7717/peerj.715

McDonald, M. A., Mesnick, S.-L., & Hildebrand, J. A. (2006). Biogeographic characterization of blue whale song worldwide.: Using song to identify populations. *Journal of Cetacean Research and Management, 8*(1), 55–65. Retrieved from https://escholarship.org/uc/item/5r16c2mz

Mellinger, D. K., & Clark, C. W. (2003). Blue whale (Balaenoptera musculus) sounds from the North Atlantic. *The Journal of the Acoustical Society of America, 114*(2), 1108–1119. https://doi.org/10.1121/1.1593066

Miller, S. A., Dykes, D. D., & Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acid Research, 16*(3).

Nei, M. (1987). *Molecular Evolutionary Genetics*: Columbia University Press.

Onoufriou, A. B., Gaggiotti, O. E., Aguilar de Soto, N., McCarthy, M. L., Morin, P. A., Rosso, M., . . . Carroll, E. L. (2022). Biogeography in the deep: Hierarchical population genomic structure of two beaked whale species. *Global Ecology and Conservation, 40*(1), e02308. https://doi.org/10.1016/j.gecco.2022.e02308

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., . . . Dalén, L. (2015). Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology : CB*, *25*(10), 1395–1400. https://doi.org/10.1016/j.cub.2015.04.007

Pampoulie, C., Gíslason, D., Ólafsdóttir, G., Chosson, V., Halldórsson, S. D., Mariani, S., . . . Víkingsson, G. A. (2021). Evidence of unidirectional hybridization and second-generation adult hybrid between the two largest animals on Earth, the fin and blue whales. *Evolutionary Applications*, *14*(2), 314–321. https://doi.org/10.1111/eva.13091

Paradis, E., & Schliep, K. (2019). ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics (Oxford, England)*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pastene, L. A., Acevedo, J., & Branch, T. A. (2019). Morphometric analysis of Chilean blue whales and implications for their taxonomy. *Marine Mammal Science*, *36*(1), 116–135. https://doi.org/10.1111/mms.12625

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., . . . Reich, D. (2012). Ancient admixture in human history. *Genetics*, *192*(3), 1065–1093. https://doi.org/10.1534/genetics.112.145037

Pike, D. G., Víkingsson, G. A., Gunnlaugsson, T., & Øien, N. (2013). A note on the distribution and abundance of blue whales (*Balaenoptera musculus*) in the Central and Northeast North Atlantic. *NAMMCO Scientific Publications*, *7*, 19. https://doi.org/10.7557/3.2703

Prasad, A., Lorenzen, E. D., & Westbury, M. V. (2022). Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Molecular Ecology Resources*, *22*(1), 45–55. https://doi.org/10.1111/1755-0998.13457

Rice, A., Širović, A., Hildebrand, J. A., Wood, M., Carbaugh-Rutland, A., & Baumann-Pickering, S. (2022). Update on frequency decline of Northeast Pacific blue whale (Balaenoptera musculus) calls. *PloS One*, *17*(4), e0266469. https://doi.org/10.1371/journal.pone.0266469

Rosel, P. E., Dizon, A. E., & Heyning, J. E. (1994). Genetic analysis of sympatric morphotypes of common dolphins (genus Delphinus). *Marine Biology*, *119*(2), 159–167. https://doi.org/10.1007/BF00349552

Rosel, P. E., Hancock-Hanser, B. L., Archer, F. I., Robertson, K. M., Martien, K. K., Leslie, M. S., . . . Taylor, B. L. (2017). Examining metrics and magnitudes of molecular genetic differentiation used to delimit cetacean subspecies based on mitochondrial DNA control region sequences. *Marine Mammal Science*, *33*(S1), 76–100. https://doi.org/10.1111/mms.12410

Sambrook, J., & Russell, D. W. (2006). Isolation of High-molecular-weight DNA from Mammalian Cells Using Formamide. *CSH Protocols*, *2006*(1). https://doi.org/10.1101/pdb.prot3225

Schiffels, S., & Wang, K. (2020). MSMC and MSMC2: The Multiple Sequentially Markovian Coalescent. *Methods in Molecular Biology (Clifton, N.J.)*, *2090*, 147–166. https://doi.org/10.1007/978-1-0716-0199-0_7

Schmidt, D., & Pool, J. (2002). The effect of population history on the distribution of the Tajima's D statistic. *Population English Edition*, 1–8.

Sears, R., & Perrin, W. F. (2009). Blue Whale: Balaenoptera musculus. In W. F. Perrin, B. G. Würsig, & J. G. M. Thewissen (Eds.), *Encyclopedia of marine mammals* (2nd ed.). Amsterdam, Boston, Mass.: Elsevier/Academic Press.

Slater, G. J., Goldbogen, J. A., & Pyenson, N. D. (2017). Independent evolution of baleen whale gigantism linked to Plio-Pleistocene ocean dynamics. *Proceedings. Biological Sciences*, *284*(1855). https://doi.org/10.1098/rspb.2017.0546

Stafford, K. M., Nieukirk, S. L., & Fox, C. G. (2001). Geographic and seasonal variation of blue whale calls in the North Pacific. *Journal of Cetacean Research and Management*, *3*(1), 65–76.

Taylor, B. L., Chivers, S. J., Larese, J., & Perrin, W. F. (2007). Generation length and percent mature estimates for IUCN assessments of cetaceans. *Administrative Report, LJ-07-01*.

Taylor, B. L., Archer, F. I., Martien, K. K., Rosel, P. E., Hancock-Hanser, B. L., Lang, A. R., . . . Baker, C. S. (2017). Guidelines and quantitative standards to improve consistency in cetacean subspecies and species delimitation relying on molecular genetic data. *Marine Mammal Science*, *33*(S1), 132–155. https://doi.org/10.1111/mms.12411

Taylor, B. L., Perrin, W. F., Reeves, R. R., Rosel, P. E., Wang, J. Y., Cipriano, F., . . . Brownell, R. L. (2017). Why we should develop guidelines and quantitative standards for using genetic data to delimit subspecies for data-poor organisms like cetaceans. *Marine Mammal Science*, *33*(S1), 12–26. https://doi.org/10.1111/mms.12413

Teixeira, J. C., & Huber, C. D. (2021). The inflated significance of neutral genetic diversity in conservation genetics. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(10). https://doi.org/10.1073/pnas.2015096118

Tønnessen, J. N., & Johnsen, A. O. (1982). *The history of modern whaling*. Berkeley: University of California Press.

Van der Valk, T., Manuel, M. de, Marques-Bonet, T., & Guschanski, K. (2019). Estimates of genetic load suggest frequent purging of deleterious alleles in small populations. *bioRxiv, 696831*. Retrieved from https://doi.org/10.1101/696831

Wolf, M., Jong, M. de, Halldórsson, S. D., Árnason, Ú., & Janke, A. (2022). Genomic Impact of Whaling in North Atlantic Fin Whales. *Molecular Biology and Evolution*, *39*(5). https://doi.org/10.1093/molbev/msac094

[dataset] Wolf, M., Jong, M. J. de, & Janke, A. (2023). Supporting Data for: Genomes of northern hemisphere and pygmy blue whales suggest isolated subspecies. *Dryad, Dataset*. Retrieved from https://doi.org/10.5061/dryad.47d7wm3jz

Data Accessibility and Benefit-Sharing

Data Accessibility

Raw sequencing reads of blue whales have been deposited at the National Center for Biotechnology Information under the BioProject PRJNA955240. Raw short read data for the single sei whale individual can be found under the BioProject PRJNA957797. Most of the code used to generate the results presented can be found on GitHub: /mag-wolf/RESEQ-to-Popanalyses/ and Zenodo: https://doi.org/XX.XXXX/zenodo.XXXXXXX. All secondary data, namely the SNP datasets are uploaded to a Dryad repository: [dataset] (Wolf, Jong, & Janke, 2023). All other data needed to evaluate the conclusions of the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper can be requested from the authors.

Benefit-Sharing

Benefits from this research accrue from providing scientific information relevant to conservation and sustainable use of biological diversity as well as by sharing of our data and results on public databases as described above.

Author contributions

M.W. and A.J. conceived and designed the study, M.W., M.J., A.L. and A.J. wrote the manuscript, M.W. made the analyses, M.J. aided with the computational analyses of the re-sequencing data.

Figures

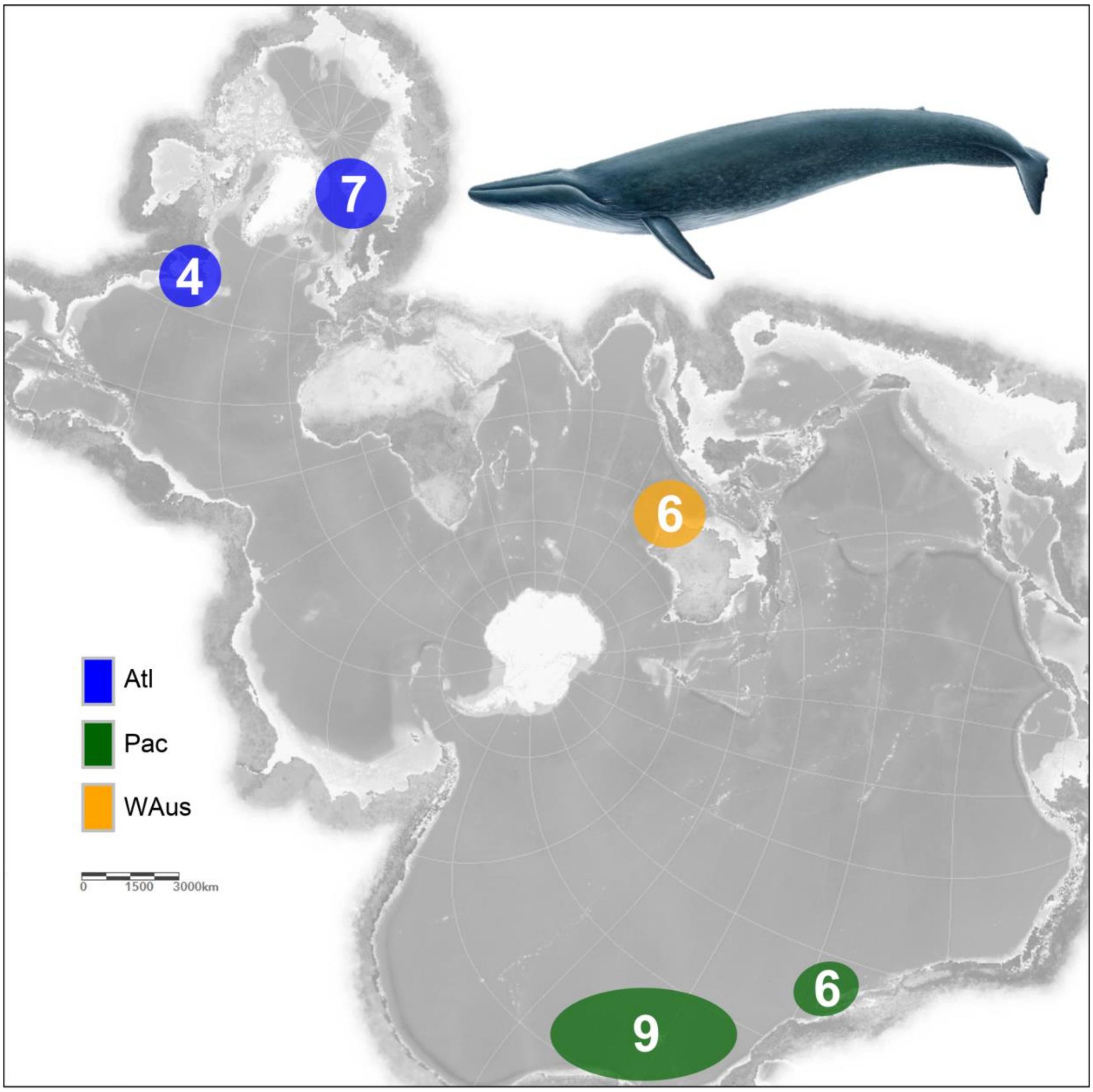

Fig 1. Geographic distribution of sampling locations of blue whale individuals analyzed in this study. Numbers indicate the number of samples, colors were assigned to the respective oceanic region: blue: North-Atlantic (**Atl**); green: North-Pacific (**Pac**); yellow: Western-Australia (**WAus**). North-Pacific and Western-Australian samples were sequenced and provided by this study. Spilhaus projection map retrieved from ArcGis. Illustration made by Jón Baldur.

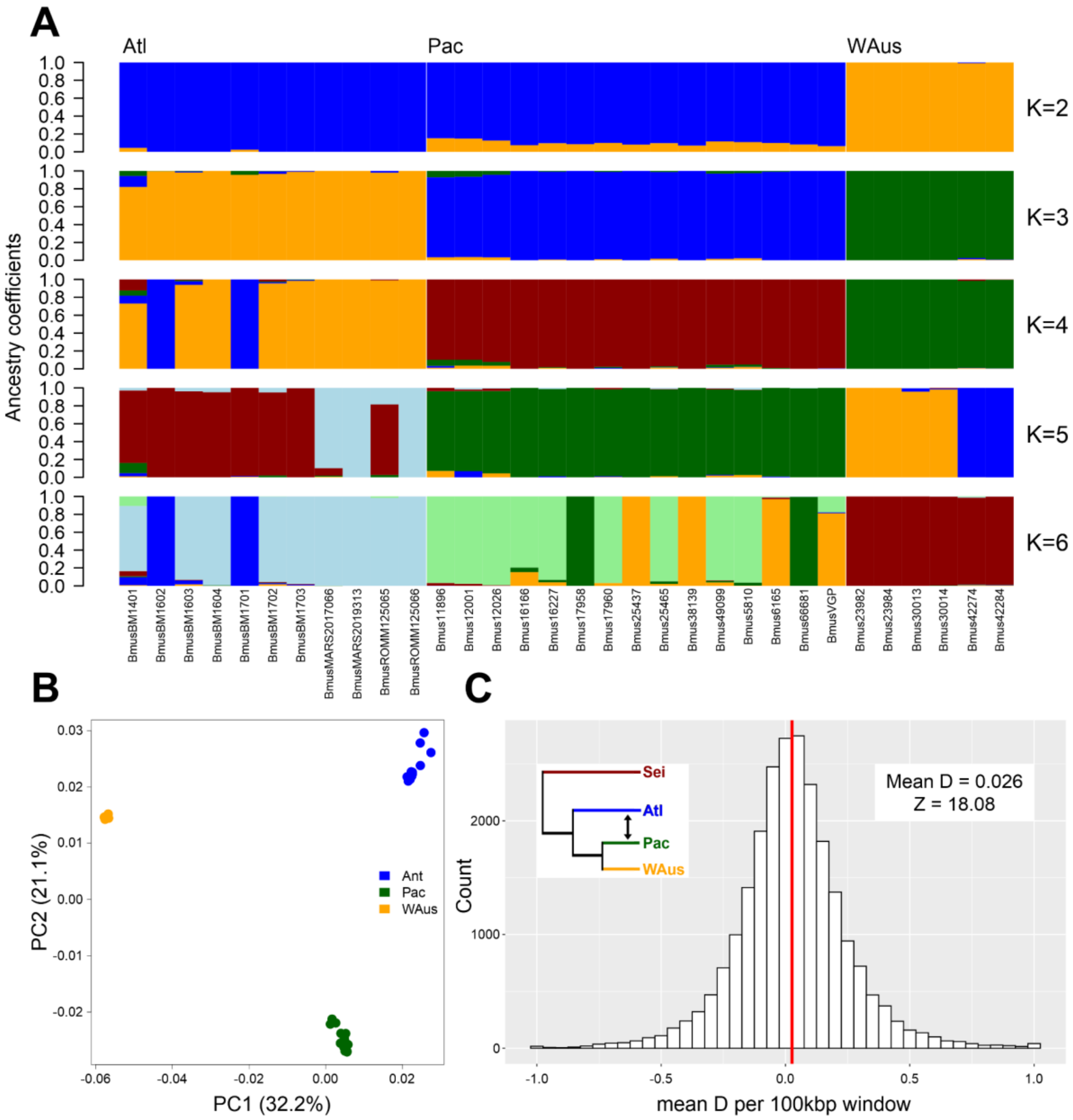

Fig. 2 Population structure and gene flow analyses in blue whales from the North-Atlantic, North-Pacific and Western-Australian waters. **A** An admixture-like test generated with the LEA R package to infer ancestry coefficients assuming 2-6 ancestral populations. Apart from the clear segmentation into the three oceanic regions, no consistent pattern was found. Signs of admixture were rare and only consistent for the Atlantic individual BM1401, which shows signs of genetic exchange with both North-Pacific and western Australia. Colors do not necessarily correspond to colors assigned to sample origin. **B** Principal Coordinates Analysis (PCoA) generated with the ape R package. All three oceanic regions were separated into distinct clusters. **C** D-statistical assessment of gene flow between North-Atlantic and North-Pacific blue whales assuming the topology presented in Fig. 2. The histogram depicts the distribution of D values over 100 kbp non-overlapping sliding windows. The mean window-based D value was slightly positive indicating gene flow between North-Atlantic and North-Pacific blue whales. A high Z score confirms the consistency of this pattern across the chromosomes.

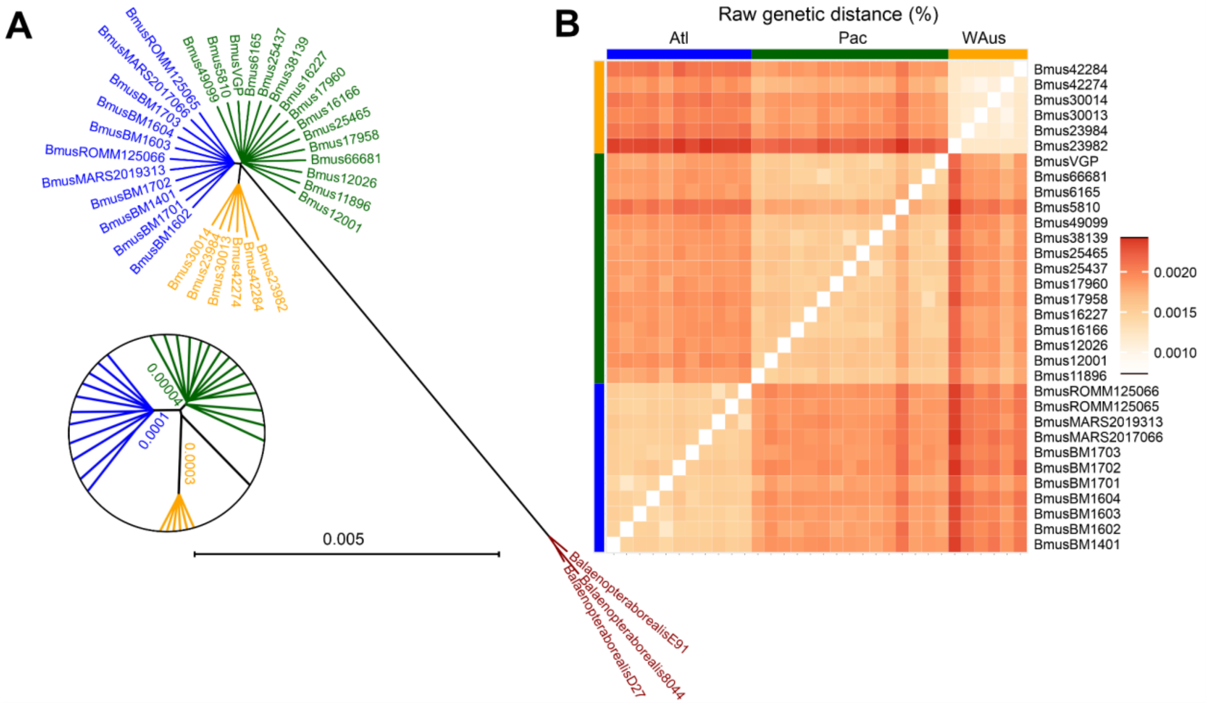

Fig. 3 Genetic distances among blue whales from different oceanic regions, including a sei whale outgroup. **A** BIONJ based phylogeny constructed with the ape v5.3 R package. All oceanic regions form distinct clusters and branch lengths indicate the highest amount of genetic distance between Western-Australian blue whales and other populations. The tree groups together North-Atlantic and North-Pacific blue whales, although to a minimal extent. **B** Heatmap representing pairwise genetic distances between blue whale individuals. The largest distance was found between North-Atlantic and Western-Australian individuals. Western-Australian individuals displayed a comparable low population-intern distance. Two outlier individuals were found with a higher genetic distance to all other individuals, namely 5810 (Pac), 23982 (WAus).

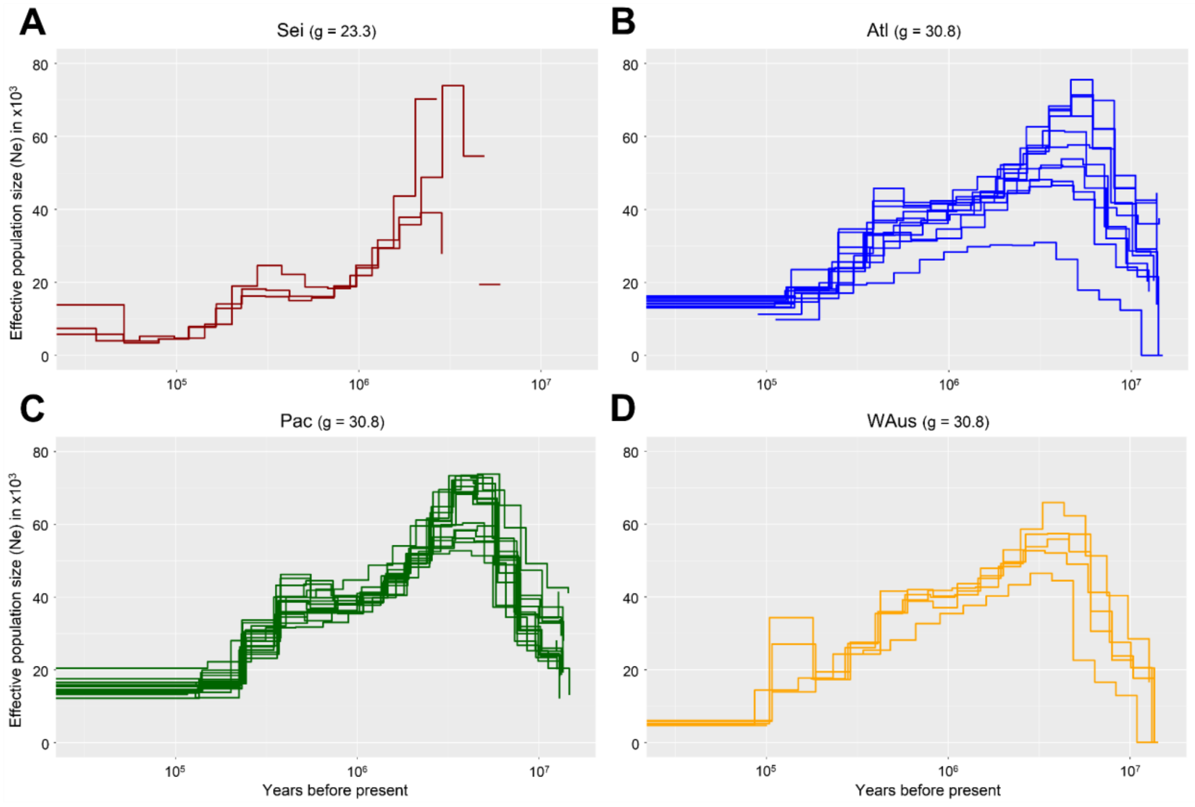

Fig. 4 Demographic changes in Ne over time in a sei whale (**A**) and different blue whale populations (**B** North-Atlantic, **C** North-Pacific, **D** Western-Australia), modeled via the MSMC2 framework. Graphs were scaled using a mutation rate of 1.38e-8 per site per generation (Árnason et al., 2018) and a generation time of 23.3 years for the sei whales and 30.8 years for the blue whale populations, respectively (Taylor, Chivers, Larese, & Perrin, 2007). y-axis depicts effective population size in x10^{3}, x-axis depicts years before present between 10^{7} and 10^{5} years ago (log10). All models indicate a peak at ~4-5 million years ago followed by a steady decline. All blue whale models show a more stable population size between 300 thousand and 1.5 million years ago. Western-Australian blue whale models also suggest a last steep population peak at around 100 thousand years ago, followed by a decline which results in a lower final population size compared to the other two populations. In general, blue whale models were too similar to infer a split between different populations. After 10 thousand years ago, all models indicate a lack of coalescence events that do not allow further assessments of demographic changes.

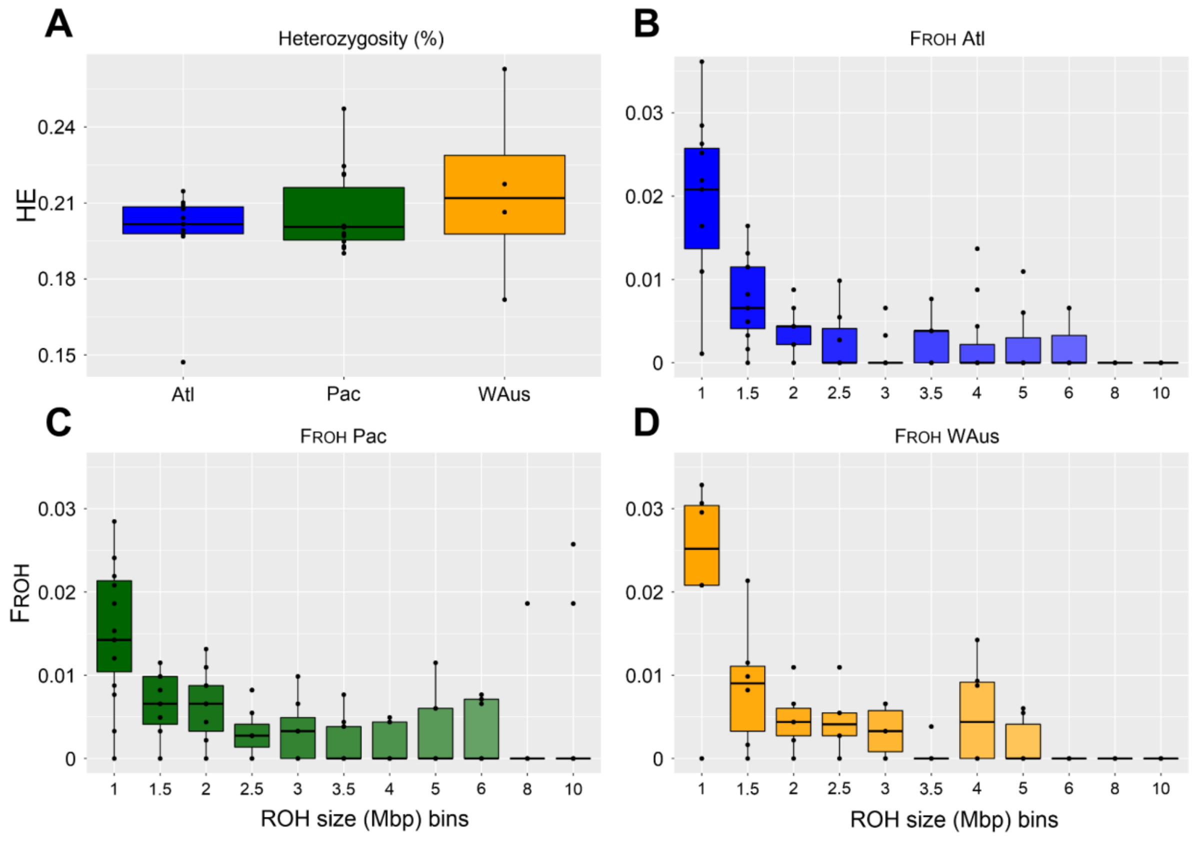

Fig. 5 Genomic diversity and inbreeding coefficients in the three inferred blue whale populations from different oceanic regions (North-Atlantic, North-Pacific and Western-Australia). **A** Genome-wide levels of heterozygosity were similar in Pacific and Atlantic individuals (Pac=0.206, Atl=0.199) and highest and most diverse in Western-Australian individuals (WAus=0.215) although non-significant. **B-D** Inbreeding coefficients based on runs of homozygosity (F_{ROH}) sorted in different size-bins (1-10 Mbp) compared over the three inferred blue whale populations. All populations featured ROHs in bins ranging from >1 to <6 Mbp. ROHs over 8 Mbp were only found in the Pacific cohort. In bins that allow for statistical comparisons, no significant differences between whale populations were found.

Tables

Table 1 Differentiation estimates inferred for 32 whole genome re-sequencing datasets, mitogenomic and mtMarker data compiled from blue whales of three different oceanic regions (North-Atlantic, North-Pacific and Western Australia), excluding outliers, as well as three Sei whale samples. The table includes pairwise comparisons of mean F_{ST} (Hudson 1992, Bhatia et al. 2013) and d_A (Nei 1987) values.

	Population	**Atl**	**Pac**	**WAus**	**Sei**	
Genomic	**Atl**	-	0.0141	0.0385	0.7974	**D_A**
	Pac	0.0619	-	0.0323	0.7977	
	WAus	0.1470	0.1247	-	0.8179	
Mitogenomic	**Atl**	-	0.0196	0.0362	3.1834	
	Pac	0.11	-	0.0151	3.1947	
	WAus	0.227	0.145	-	3.2313	
mtMarker	**Atl**	-	0.0856	0.4312	3.3754	
	Pac	0.0709	-	0.2398	3.9429	
	WAus	0.3129	0.2648	-	4.4903	
		F_{ST}				**Metrics**

Table 2 Diversity statistics inferred for the three blue whale populations North-Atlantic, North-Pacific and Western Australia. The table includes statistics for mean genome-wide heterozygosity in % (He), nucleotide diversity in % (π), Tajima's D, Wattersons Θ and inbreeding coefficients based on ROH of 1 Mbp or longer.

Metrics	**Population**		
	Atl	**Pac**	**WAus**
Mean He	0.199	0.206	0.215
Nuc. div (π)	0.199	0.204	0.212
Tajima's D	0.882	0.594	0.857
Watterson Θ	121.1	121.3	122.4
F_{ROH} (>1 Mbp)	0.043	0.052	0.048

Supplementary Materials for

Genomes of northern hemisphere and pygmy blue whales suggest isolated subspecies.

Magnus Wolf, Menno J. de Jong, Axel Janke

*Corresponding author: Magnus Wolf; Email: Magnus.Wolf@senckenberg.de

This PDF file includes:

Figs. S1 to S5

Tables S1 to S4

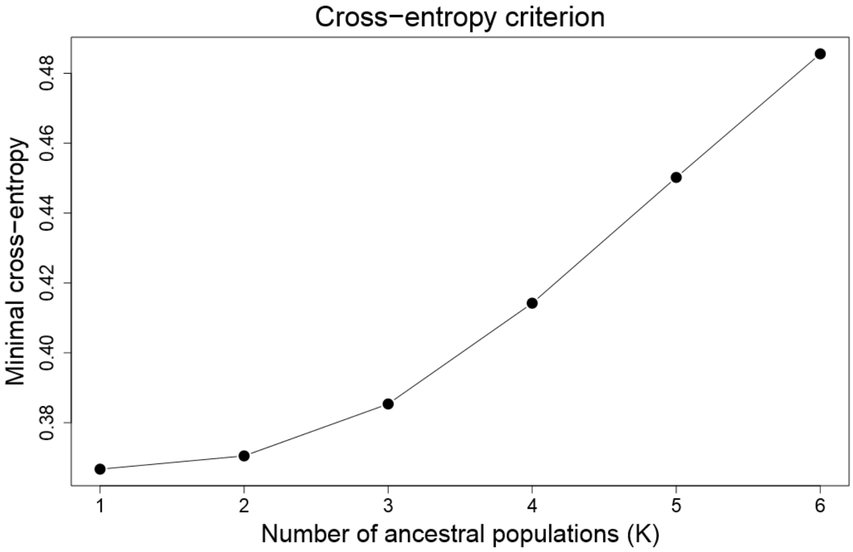

Fig. S1 Determination of the optimal K value for the admixture-like analysis performed with the LEA-2.4.0 package (Frichot et al. 2015). Shown are minimal cross-entropy scores generated with LEA's "snmf' function.

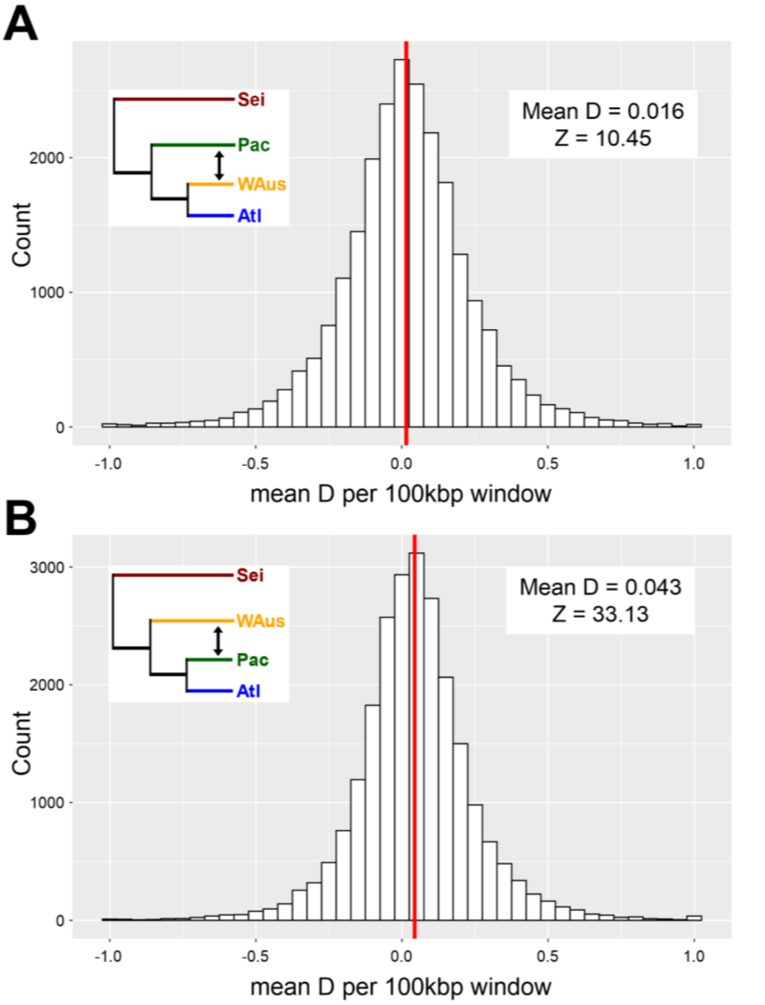

Fig. S2 D-statistic assessment of gene flow between blue whale populations testing alternative topologies. **A** Test of gene flow between the Pacific and the West-Australian population assuming the Pacific population to be the introgressor. **B** Test of gene flow between the West-Australian and the Pacific population assuming the West-Australian population to be the introgressor.

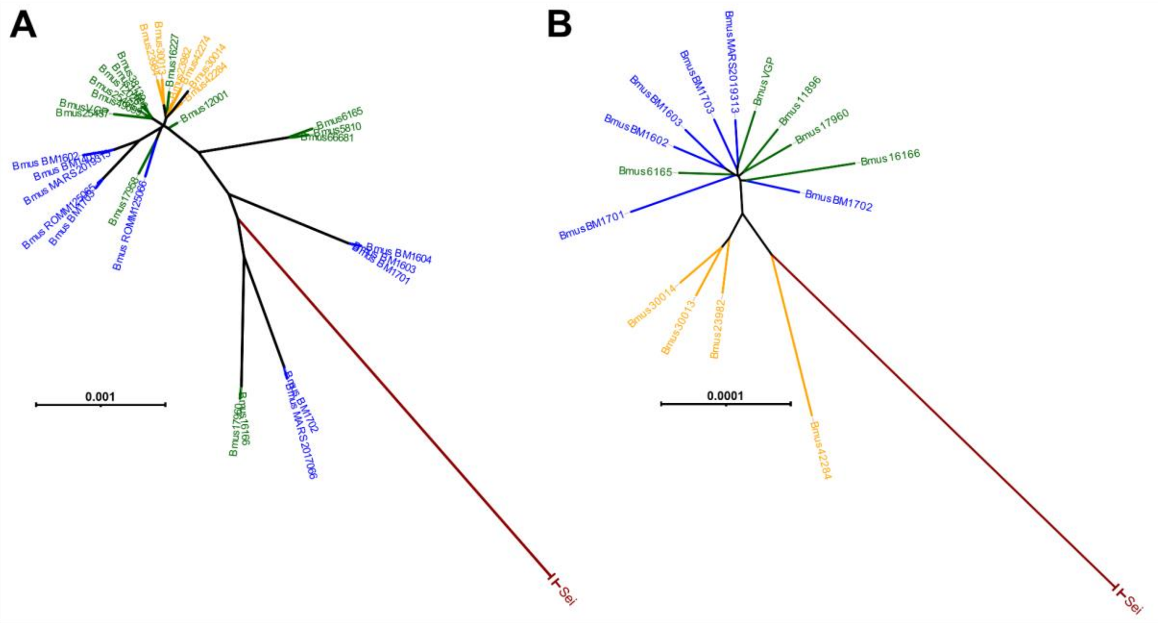

Fig. S3 Unrooted distance based BIONJ trees retrieved from haploid data using the ape v5.3 R package. **A** Mitochondrial tree. **B** Y-chromosomal tree.

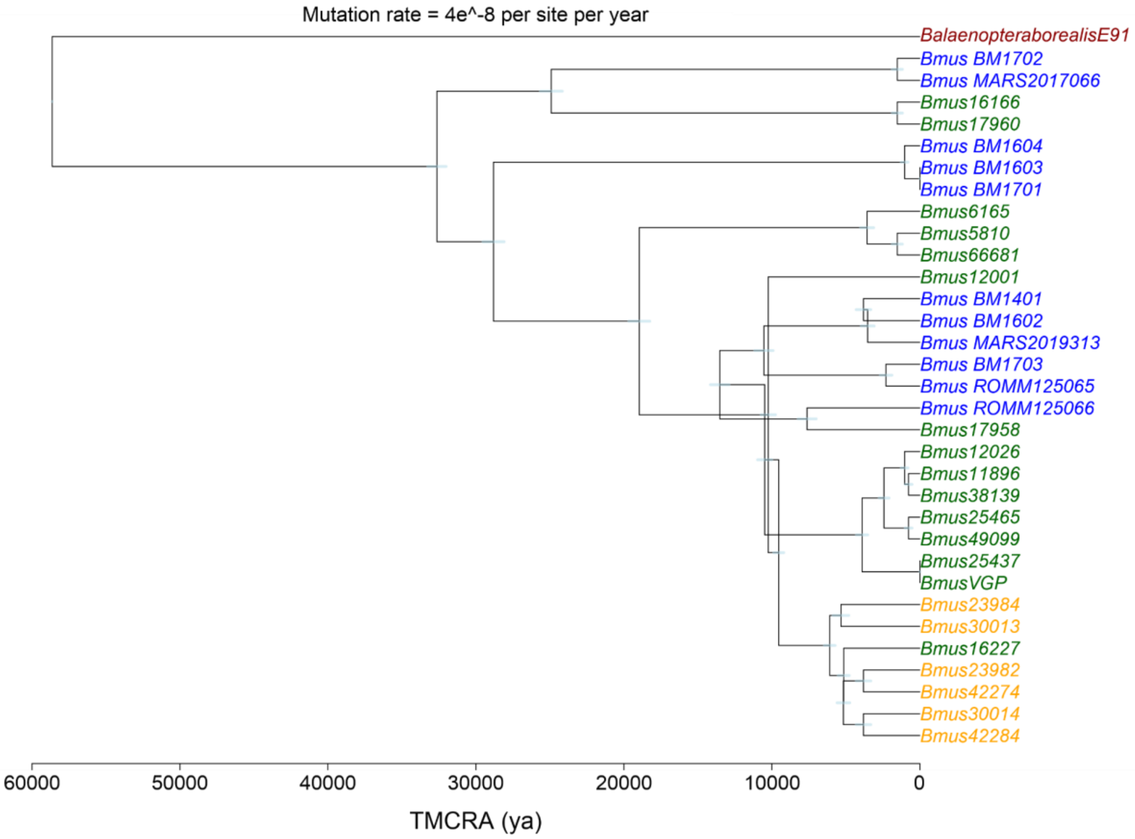

Fig. S4 Dated phylogram created with SambaR's "popprtree()" function using the mitochondrial based distances and an mutation rate of 4.0 × 10^{-8} bp^{-1} year^{-1} (per site per year) (Brüniche-Olsen et al. 2021). Although multiple nothern-hemisphere individuals fall out of the main clusters, a main split was found to be at ~10-15 kya.

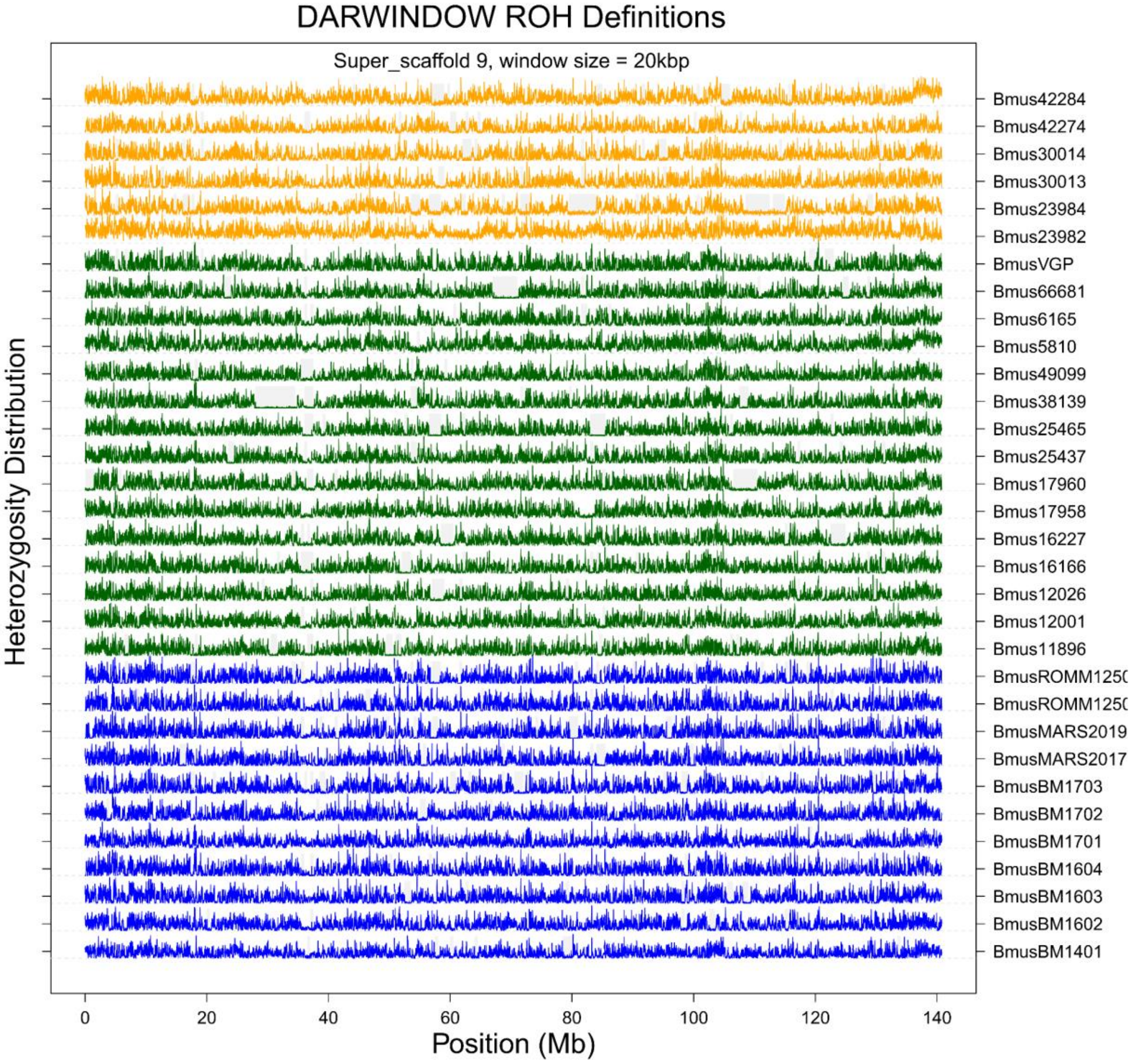

Fig. S5 Heterozygosity distribution across the randomly chosen Super_scaffold_9 depicting 20 kbp long sliding windows. Runs of homozygosity found by DARWINDOW were highlighted in gray and fit the visible drops in heterozygosity.

Table S1 General information for samples sequenced in this study. Provided are field-IDs and source codes from the SWFSC collection as well as sampling date, location and sex. For the single Sei whale, a lab ID is provided instead.

SWFSC ID	Field ID	Source	Year	Month	Day	Latitude	Longitude	Sex
5810	DSJ960728.02	BIOPSY-SWFSC-WHAPSI	1996	7	28	34,15	-120,283333	F
6165	DSJ960908.02	BIOPSY-SWFSC-ORCAWA	1996	9	8	33,85	-121,433333	M
11896	END981127.02	BIOPSY-SWFSC-SPAM	1998	11	27	-7,133333	-80,583333	M
12001	DSJ981106.01	BIOPSY-SWFSC-SPAM	1998	11	6	-1,3	-92,316666	F
12026	DSJ981109.01	BIOPSY-SWFSC-SPAM	1998	11	9	-1,833333	-86,083333	F
16166	DSJ991124.05	BIOPSY-SWFSC-STAR99	1999	11	24	8,35	-93,95	M
16227	MAC991016.04	BIOPSY-SWFSC-STAR99	1999	10	16	9,883333	-96,016666	F
17958	DSJ000730.04	BIOPSY-SWFSC-STAR00	2000	7	30	29,716666	-115,816666	F
17960	DSJ000730.06	BIOPSY-SWFSC-STAR00	2000	7	30	29,716666	-115,816666	M
25437	DSJ010921.12	BIOPSY-SWFSC-ORCAWA	2001	9	21	43,5	-124,45	F
25465	DSJ010930.12	BIOPSY-SWFSC-ORCAWA	2001	9	30	34,65	-121,05	F
38139	DSJ030921.04	BIOPSY-SWFSC-STAR03	2003	9	21	13,166666	-99,05	F
49099	HYDE051013.01	BIOPSY-SWFSC-HYDE	2005	10	13	32,85	-117,316666	-
66681	DSJ060730.11	BIOPSY-SWFSC-STAR06	2006	7	30	30,366666	-116,433333	F
23982	BMUS000328.01	AUSTRALIA-WAM-BANNIS	2000	3	28	-32,05	115,05	M
23984	BMUS000229.01	AUSTRALIA-WAM-BANNIS	2000	2	29	NULL	NULL	F
30013	BM020130.01	AUSTRALIA-WAM-BANNIS	2002	1	30	-31,916666	114,966666	M
30014	BM020319.01	AUSTRALIA-WAM-BANNIS	2002	3	19	-31,916666	115	M
42274	BMUS-040207-6	AUSTRALIA-WAM-BANNIS	2004	2	7	-31,883333	115,033333	F
42284	BMUS-040321-16	AUSTRALIA-WAM-BANNIS	2004	3	21	-31,016666	115,166666	M

Lab ID	Source	Year	Month	Day	Latitude	Longitude	Sex
8044	CELL-CULTURE, ICELAND	1986	n.a.	n.a.	n.a.	n.a.	M

Table S2 General information for samples retrieved from NCBI sequence read archive (SRA) repository. Samples are sorted by the BioProject. Provided are the sample ID provided by the submitting authors, biosample and SRA accession numbers as well as known location and sampling years.

BioProject: PRJNA704862				
ID	**Biosample**	**SRA**	**Location**	**Year**
BM1703	SAMN18057126	SRR14357022	79.92 N 14.45 E	2017
BM1702	SAMN18057125	SRR14357025	78.73 N 9.26 E	2017
BM1701	SAMN18057124	SRR14357028	79.77 N 9.47 E	2017
BM1604	SAMN18057123	SRR14357029	78.45 N 12.15 E	2016
BM1603	SAMN18057122	SRR14357030	78.33 N 12.13 E	2016
BM1602	SAMN18057121	SRR14357032	78.33 N 12.13 E	2016
BM1401	SAMN18057120	SRR14357033	Svalbard	2014
MARS2019313	SAMN18057109	SRR14357010	Canada: Sutherlands Cove, Cape Breton Island, Nova Scotia	2019
MARS2017066	SAMN18057108	SRR14357011	Canada: Ragged Harbour, Nova Scotia	2017
ROMM125065	SAMN18057107	SRR14357035	Canada: Trout River, Newfoundland and Labrador	2014
ROMM125066	SAMN18057106	SRR14467046	Canada: Rocky Harbour, Newfoundland and Labrador	2014
BioProject: PRJNA389516				
E91	SAMN07201757	SRS2268081	Iceland	-
D27	SAMN07201756	SRS2268080	Iceland	-

Table S3 Sequencing and mapping statistics for all sequenced and publicly available samples. Most statistics were collected with QUALIMAP v.2.2.2 (Okonechnikov et al. 2015). Provided are: the sample ID, the retrieved number of reads, the percentage of mapped reads, the estimated number of duplicated reads, mean insert sizes, mean mapping quality, the general error rate, the mean coverage, the coverage standard deviation, the number of retrieved SNPs after individual variant calling and filtering as described in the general method section.

ID	No. reads	map. (%)	No. dup.	insert size	map. qual.	error rate	cov.	std cov.	No. Genotypes
8044	1002713262	67.31	333507078	325498.8	52.3	0.0255	36.6	1097.4	1074255177
D27	314819542	58.62	51853177	328485.2	52.3	0.0226	6.5	165.9	1185713484
E91	311030310	56.23	50049832	335987.5	52.3	0.0230	6.1	168.3	1166774568
11896	299065816	73.36	94776896	313260.4	52.8	0.0160	11.7	406.3	1120026172
12001	310066206	72.84	98107266	323390.0	52.7	0.0164	12.0	416.9	1135367675
12026	313119553	72.86	99054969	306384.3	52.8	0.0162	12.2	419.8	1136632704
16166	332865840	72.97	105526910	314721.1	52.9	0.0156	13.0	451.9	1144583041
16227	342747354	72.58	106550949	318051.2	52.7	0.0157	13.3	427.2	1162833259
17958	319352518	74.46	105301016	310535.6	52.6	0.0161	12.5	415.9	1145284797
17960	303884812	73.44	96515165	313731.3	52.8	0.0168	11.8	398.8	1128455627
25437	309692751	72.81	96265851	303854.6	52.9	0.0155	12.1	410.7	1137371157
25465	314528954	73.08	97798675	315387.5	52.8	0.0158	12.3	403.0	1145343139
38139	308440586	73.27	97810167	303297.7	52.6	0.0157	12.0	383.7	1204192598
49099	301783257	77.06	114235174	322581.2	52.5	0.0149	12.0	450.7	1019490010
5810	256981584	83.97	111269399	277750.1	51.6	0.0158	10.2	288.6	1145150464
6165	303229759	75.8	89779560	277088.6	53.5	0.0133	12.3	369.9	1137898476
66681	295380727	74.7	101034474	335820.6	52.6	0.0158	11.6	420.2	1091480040
VGP	792170642	82.22	358076506	149926.2	53.6	0.0079	27.3	443.7	1192022008
23982	326212734	77.61	116761510	275044.8	52.6	0.0178	13.0	419.9	1127080068
23984	297223179	73.21	96789490	308749.2	52.5	0.0167	11.4	382.8	1193112097
30013	356320848	73.33	114376438	307276.4	52.8	0.0157	13.8	434.0	1161396329
30014	332432229	74.39	106811158	295282.8	52.8	0.0157	13.0	400.6	1148686642
42274	202758526	70.7	72124619	367694.8	51.1	0.0173	7.1	278.2	1131981618
42284	283704417	74.82	98537906	284394.0	52.2	0.0170	10.8	341.8	1158089121
BM1401	371281836	75.08	116530598	317297.2	52.9	0.0174	15.1	471.0	1185656170
BM1602	267204560	72.94	80863656	323908.1	52.9	0.0186	10.6	358.3	1181598689
BM1603	381649402	75.19	121328515	312444.8	53.0	0.0175	15.5	483.7	1187232718

BM1604	789197681	74.7	272581539	335318.0	53.1	0.0170	31.8	932.5	1198204665
BM1701	164172844	73.71	48123384	321394.8	52.9	0.0190	6.5	227.6	1055275853
BM1702	348980624	75.49	112019450	323874.9	52.9	0.0176	14.2	478.2	1169554630
BM1703	373832583	75.84	118893261	312347.6	53.0	0.0173	15.3	470.9	1185651257
MARS2017066	628020829	68.3	197331572	286476.8	52.8	0.0201	23.0	717.3	1129916638
MARS2019313	718789733	72.17	240853247	338577.0	53.1	0.0181	28.3	894.0	1175537818
ROMM125065	792665299	73.04	280504311	341357.8	52.8	0.0161	31.5	992.9	1191966717
ROMM125066	805021222	72.37	272981880	347630.4	53.0	0.0157	31.9	986.5	1195497540

Table S4 F-statistic inference for f2 and f3 statistics calculated after Patterson et al. 2012. **A** f2 matrix providing pairwise values. **B** f3 table providing triplet values between the tested population (color highlighted) and the two assumed ancestral populations. Table also includes standard errors, Z-scores and p-values.

A **f2 Statistics** (after Patterson et al. 2012)

f2	**Atl**	**Pac**	**WAus**	**Sei**
Atl	0	0.020	0.051	0.543
Pac	0.020	0	0.043	0.543
WAus	0.051	0.043	0	0.557
Sei	0.543	0.543	0.557	0

B **f3 Statistics** (after Patterson et al. 2012)

Test Pop1	**Anc. Pop2**	**Anc. Pop3**	**f3**	**SE**	**Z-score**	**P**
Atl	Pac	WAus	0.0147	0.00015	97.48	0
Pac	WAus	Atl	0.0057	0.00013	45.72	0
WAus	Atl	Pac	0.0372	0.00027	135.48	0
Sei	Atl	Pac	0.5384	0.00072	774.44	0
Sei	WAus	Atl	0.5363	0.00073	735.09	0
Sei	WAus	Pac	0.5384	0.00073	741.82	0
Atl	Sei	Pac	0.0053	0.00008	61.31	0
Atl	Sei	WAus	0.0073	0.00013	57.52	0
Pac	Sei	Atl	0.0048	0.00008	58.64	0
Pac	Sei	WAus	0.005	0.00012	41.54	0
WAus	Sei	Atl	0.0189	0.00017	114.57	0
WAus	Sei	Pac	0.0168	0.00015	109.61	0

7. Acknowledgements

Within the last 3.5 years, many people contributed to my PhD, and they all deserve my deepest gratitude and being mentioned here, because I would not have made it so far without them.

I would like to thank **Prof. Dr. Axel Janke**, for providing me with the opportunity to work on these iconic animals, for the phantastic scientific environment he created in the TBG and BiK-F and his incredible efforts in supervising me as well as commenting and contributing to my manuscripts.

I would like to express my gratitude to **Prof. Dr. Ingo Ebersberger** for co-supervising this thesis.

I would like to thank **Dr. Menno J. de Jong** for his constant input and helpful discussions on population genetics and conservation genetics. It cannot be emphasized enough how much I learned from him.

I would like to thank all my colleagues at the TBG and BiK-F that also provided important insights, helped me with lab work when I was not able to, helped with paperwork or as IT support. This includes but is not restricted to: **Dr. Sven Winter, Konstantin Zapf, Jordi di Raad, Raphael Coimbra, Marcel Nebenfuehr, David Prochotta, Dr. Bruno Ferrette, Dr. Maria Nilsson, Dr. Carola Greve, Andrea Kolb, Ricarda Prinz, Damian Baranski, Alexander Ben Hamadou, Barbara Steigler, Dr. Barbara Feldmeyer, Prof. Dr. Markus Pfenninger, Dr. Tilman Schell, Prof. Dr. Michael Hiller, Deepak Kumar Gupta, Christoph Sinai,** and **Prof. Dr. Úlfur Árnason**.

Apart of the scientific support, many people contributed to this PhD by helping me in other ways.

I would like to thank my wife, **Anna Wolf** for her help in keeping me motivated and focused and the sacrifices she made along the way that allowed me to pursue my PhD.

I would like to thank my little son **Matheo**, who's twirly, hyper-active and cute personality always made my day.

I would like to thank my parents **Detlef** and **Beatrix**, my brother **Niklas**, and my family in law, namely **Claudia, Lara, Kathi, Thomas,** and **Clara**, for their constant support, the helpful talks and taking care of my child when I wasn't able to.

Eventually, I would like to thank my friends from university times, church, and martial arts for providing a little bit of actual life in my non-existing work-life balance.

8. Danksagung

Innerhalb der letzten 3,5 Jahre haben viele Leute zu meinem Doktor beigetragen und alle verdienen meinen Dank und eine Erwähnung hier, da ich ohne sie nie so weit gekommen wäre.

Danken möchte ich **Prof. Dr. Axel Janke**, für die Chance, die er mir gegeben hat, an diesen faszinierenden Tieren zu arbeiten, für die fantastische wissenschaftliche Umgebung, die er im TBG und BiK-F geschaffen hat, und für seinen großartigen Einsatz als mein Betreuer und während meiner Schreibphasen.

Danken möchte ich auch **Prof. Dr. Ingo Ebersberger** für seine Hilfe als mein Zweitbetreuer.

Danken möchte ich **Dr. Menno J. de Jong**, für seinen konstanten Input und die hilfreichen Gespräche über Populations- und Naturschutzgenetik. Sein Beitrag zu meinem Verständnis dieser Themen kann kaum überschätzt werden.

Danken möchte ich auch all meinen Kollegen am TBG und am BiK-F für ebenfalls wichtige Erkenntnisse, für ihre Hilfe im Labor, wenn ich nicht konnte oder für die Hilfe mit Unterlagen und den Computer-Servern. Dieser Dank gilt besonders für **Dr. Sven Winter, Konstantin Zapf, Jordi di Raad, Raphael Coimbra, Marcel Nebenfuehr, David Prochotta, Dr. Bruno Ferrette, Dr. Maria Nilsson, Dr. Carola Greve, Andrea Kolb, Ricarda Prinz, Damian Baranski, Alexander Ben Hamadou, Barbara Steigler, Dr. Barbara Feldmeyer, Prof. Dr. Markus Pfenninger, Dr. Tilman Schell, Prof. Dr. Michael Hiller, Deepak Kumar Gupta, Christoph Sinai** und **Prof. Dr. Úlfur Árnason**.

Abgesehen von fachlicher Hilfe möchte ich auch noch vielen Leuten danken, die mir auf andere Weise geholfen haben.

Danken möchte ich meiner Frau, **Anna Wolf**, für ihre Hilfe daran, mich motiviert und fokussiert zu halten und für die vielen Male, die sie zurückstecken musste, damit ich an meinem Doktor arbeiten kann.

Danken möchte ich auch meinem kleinen Sohn **Matheo**, dessen quirlige, hyperaktive und süße Art immer wieder meinen Tag gerettet hat.

Danken möchte ich meinen Eltern, **Detlef** und **Beatrix**, meinem Bruder **Niklas,** sowie der Familie meiner Frau, besonders **Claudia, Lara, Kathi, Thomas** und **Clara** für ihre unermüdliche Unterstützung, für die hilfreichen Gespräche und für die vielen Male, an denen sie meinen Sohn betreut haben, wenn ich es nicht konnte.

Zuletzt möchte ich auch meinen Freunden aus Uni-Zeiten, meiner Kirche und Freunden aus Kampfsport-Zeiten danken, dafür dass sie mir ein bisschen „Life“ in meiner fehlenden Work-Life-Balance gegeben haben.