Science Advances

AAAS

advances.sciencemag.org/cgi/content/full/4/4/eaap9873/DC1

Supplementary Materials for

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

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

Published 4 April 2018, *Sci. Adv.* **4**, eaap9873 (2018)
DOI: 10.1126/sciadv.aap9873

The PDF file includes:

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

- table S8. Calibration points used for the divergence time tree, node age estimates in million years ago, and references.
- table S9. Divergence time estimates for Artiodactyla and Cetacea for nodes in the divergence time tree (Fig. 5).

Other Supplementary Material for this manuscript includes the following:
(available at advances.sciencemag.org/cgi/content/full/4/4/eaap9873/DC1)

- data S1 (Microsoft Excel format). *D* statistics results.
- data S2 (Microsoft Excel format). D_{FOIL} results.

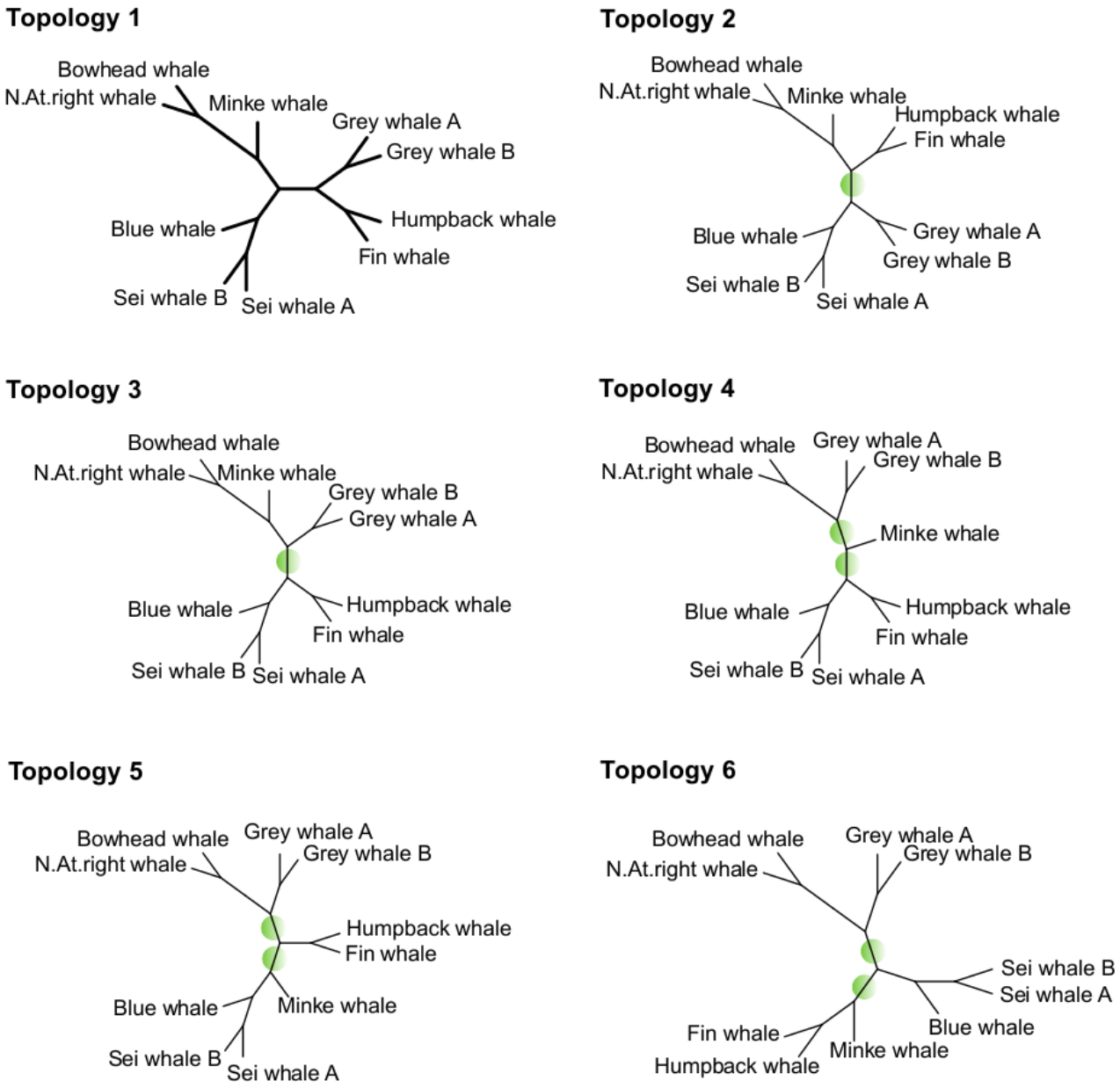

fig. S1. Possible tree topologies for baleen whales that were evaluated by the AU test.

Branch swaps that are made relative the species tree (topology 1) are marked by green dots.

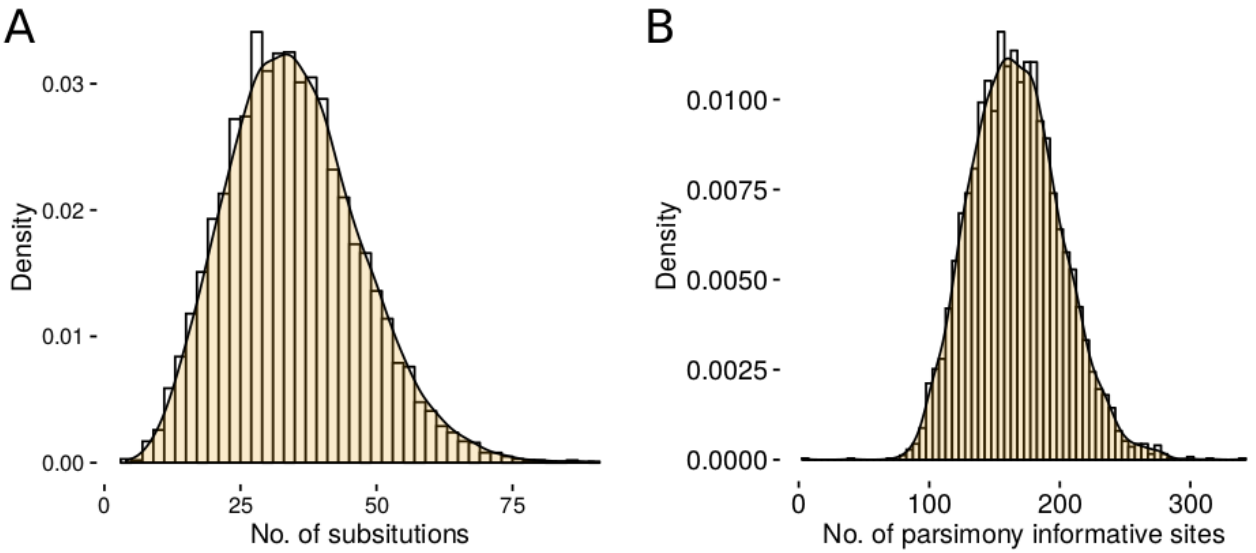

fig. S2. Phylogenetic content of GFs. (**A**) Phylogenetic content of 5,000 randomly sampled 10 kb genome fragments shown as distribution of absolute genetic distance between the North Atlantic right whale and bowhead whale, two closest related species in the taxon sampling. (**B**) Distribution of parsimony informative sites among 5,000 10 kb alignments of the baleen whales.

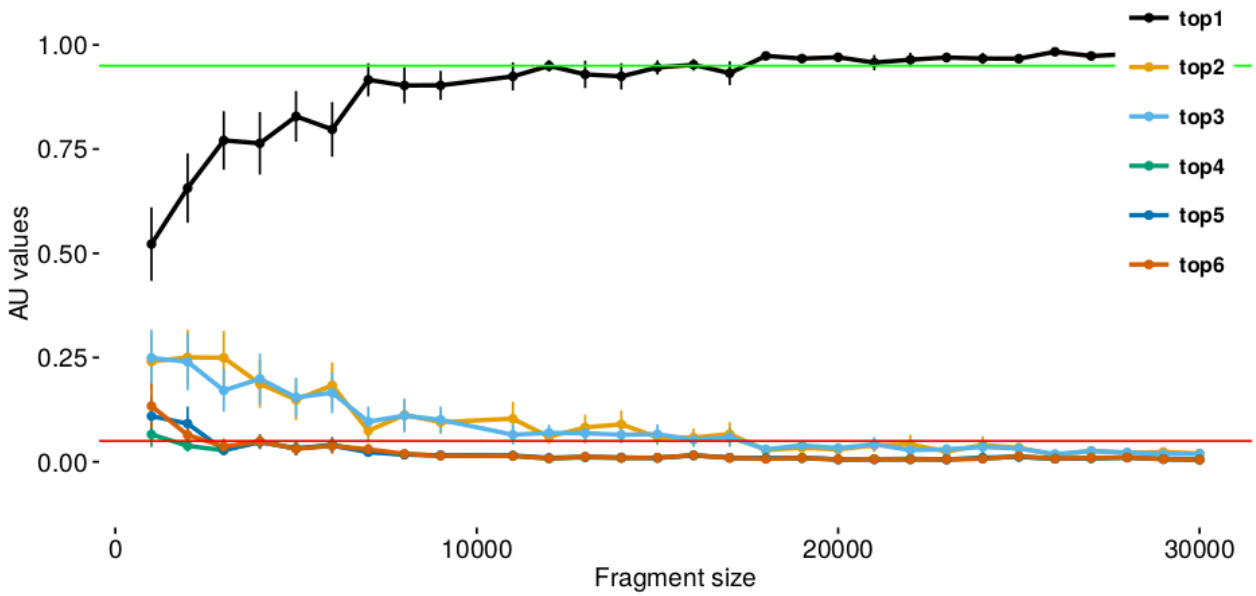

fig. S3. AU test for increasing GF sizes. GF between 1 and 100 kb were simulated using the presumed species tree and real data as input sequence. The AU test evaluates if the data rejects (pAU <0.05, red line) a phylogenetic hypothesis. The six different topologies are shown in fig. S1.

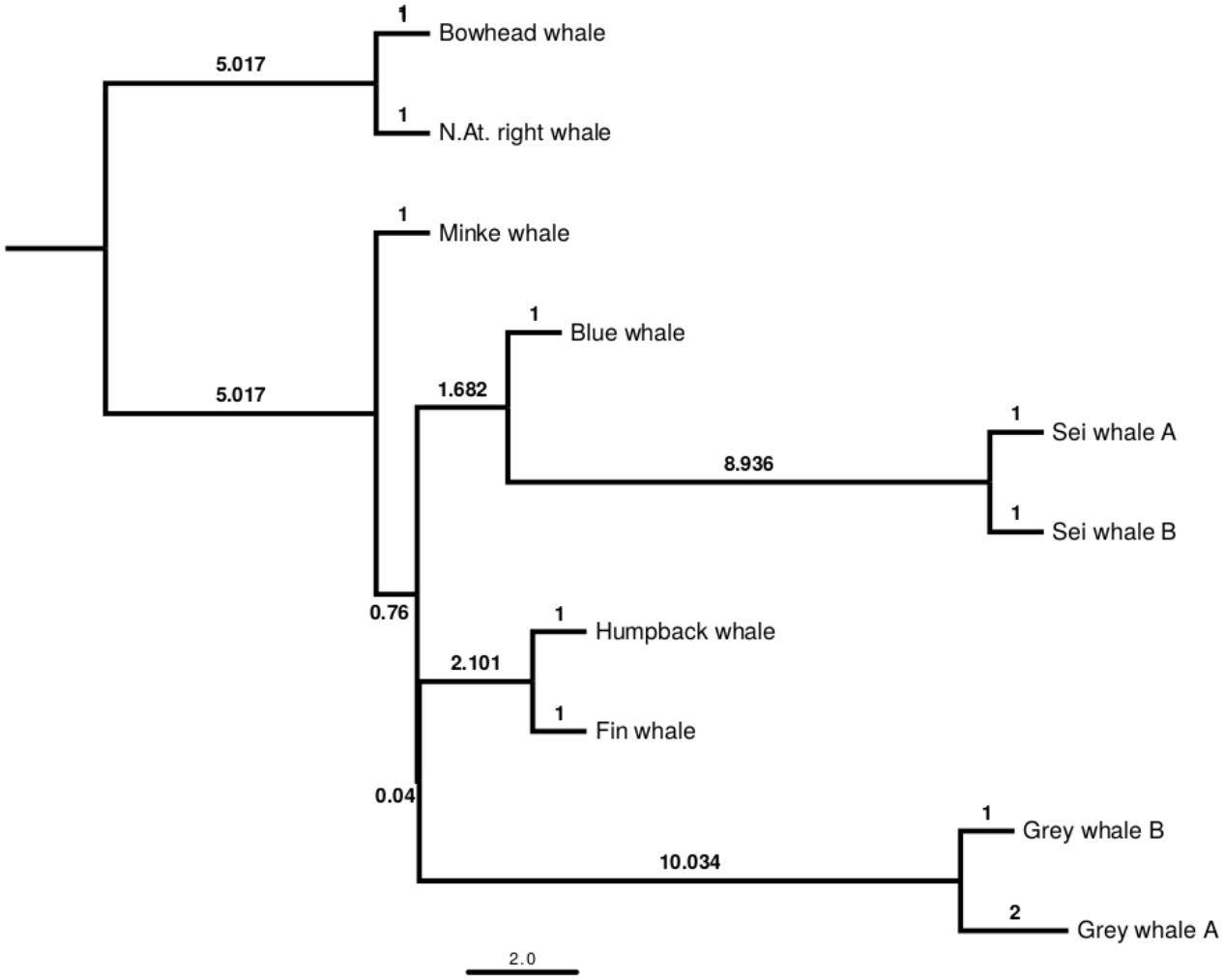

fig. S4. MSC-based species trees generated by ASTRAL using 34,192 GFs, with each GF being 20 kbp long. The tree was rooted with Bowhead and North Atlantic right whale. Branch lengths are given in coalescent units and are an indicator of gene-tree discordance, i.e. shorter branches indicate higher gene tree discordance. All branches received unanimous support in the ASTRAL analysis (posterior probability 1.0). N. At. = North Atlantic.

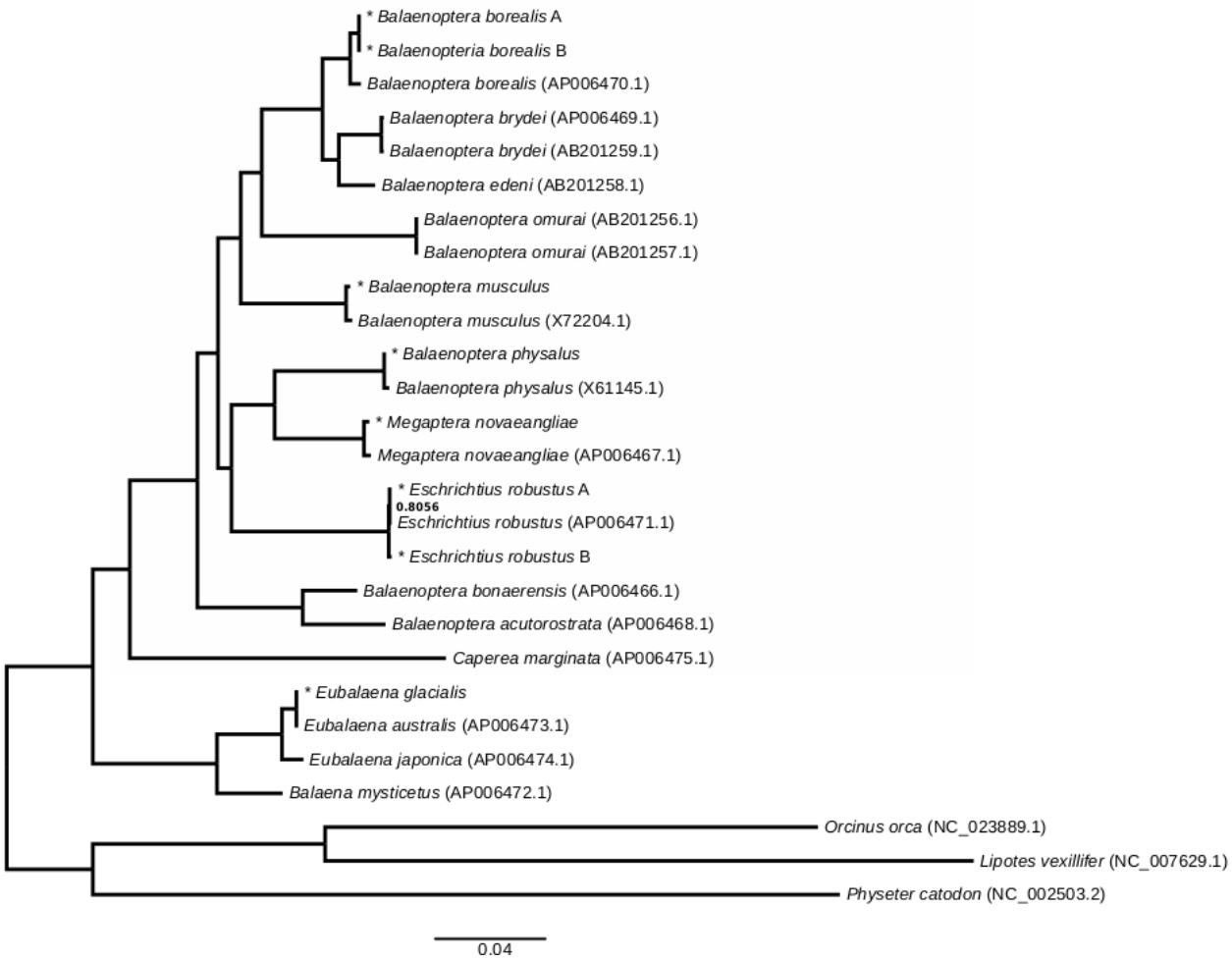

fig. S5. Phylogenetic tree from mitochondrial genomes for baleen whales. New sequences are marked with an asterisk. Accession numbers of published sequences are given in parentheses. Posterior probabilities are given at nodes if not 1.0.

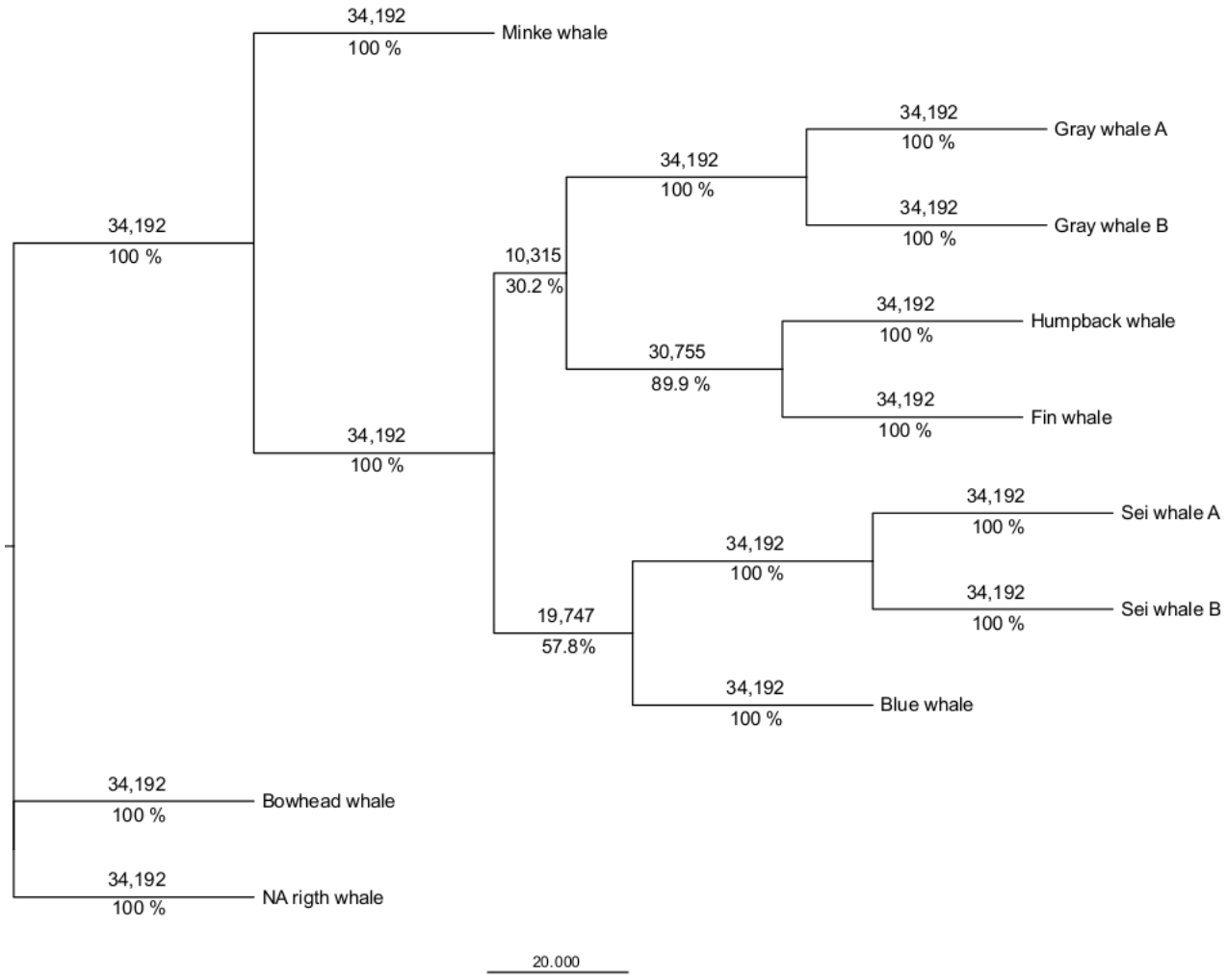

fig. S6. A majority-rule consensus tree from 34,192 individual GF ML trees (table S6) calculated with the program CONSENSE of the PHYLIP package. The topology is congruent to the coalescent species tree. Number above each branches indicate the absolute number of splits found in 34,192 individual GF trees, the number below shows the percentage values.

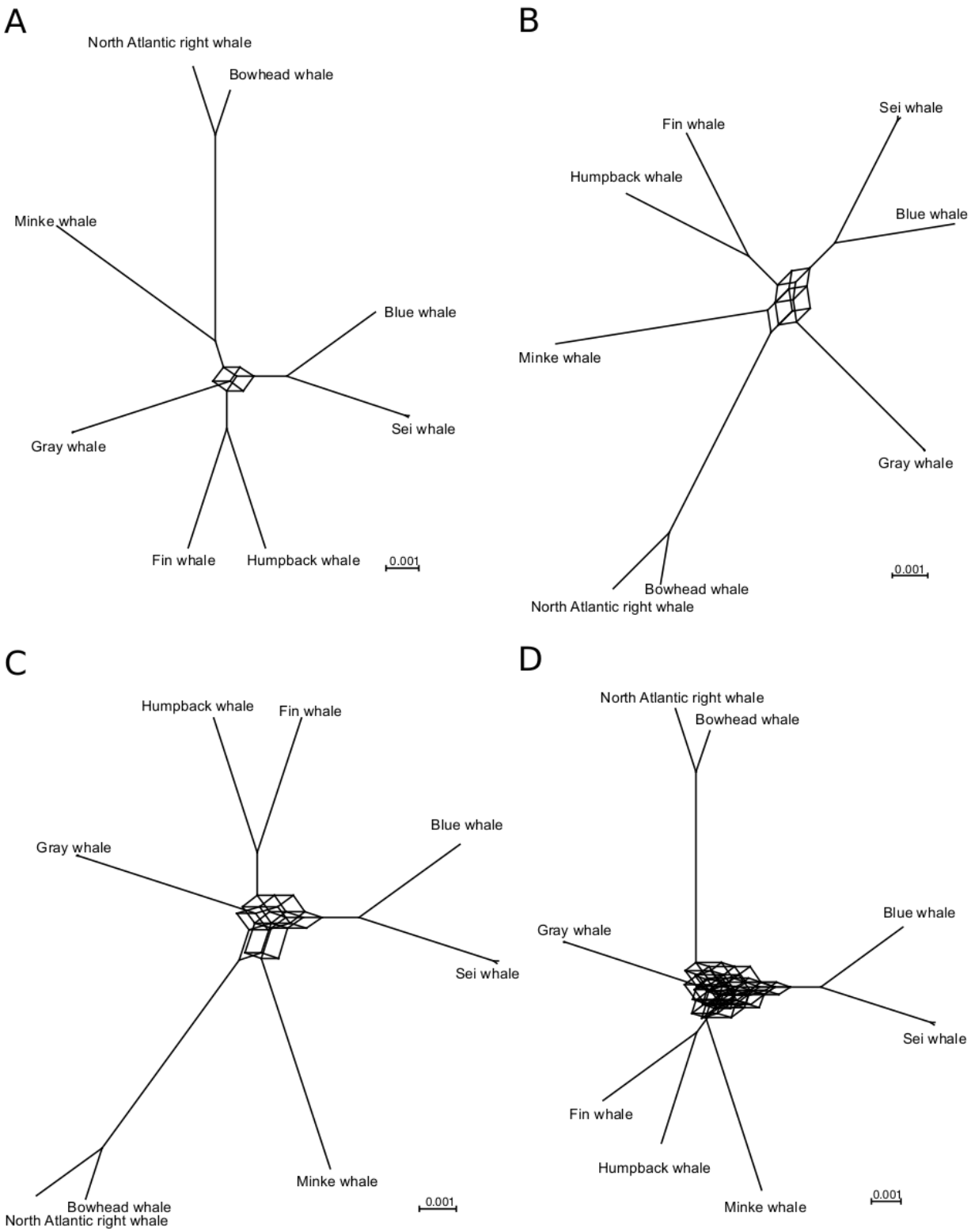

fig. S7. Consensus networks for baleen whales from 34,192 gene trees (10-kbp GF) at different minimum thresholds of gene trees to form an edge. (**A**) 14% threshold. (**B**) 11% threshold (**C**) 7% threshold, (**D**) 5% threshold.

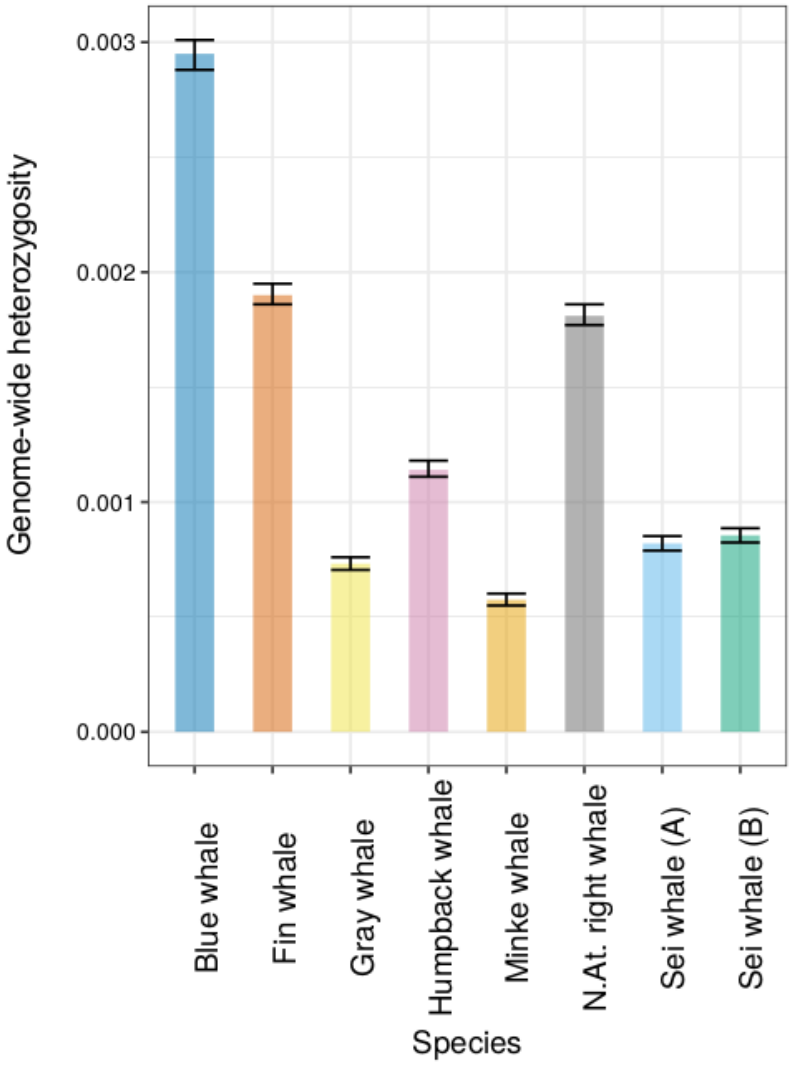

fig. S8. ML estimates of genome-wide heterozygosity estimated with mlRho. Error-bars indicate confidence intervals.

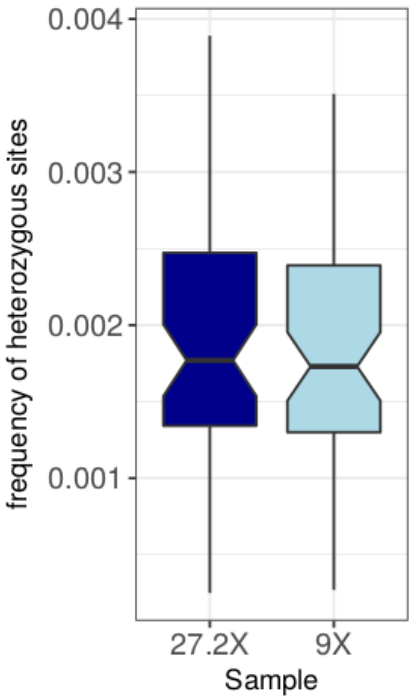

fig. S9. Blue whale heterozygosity for different sequencing depth. The sequencing depth does not affect the frequency of heterozygous sites. Heterozygosity was estimated in 100 kb windows with a total of 5.8 Mb.

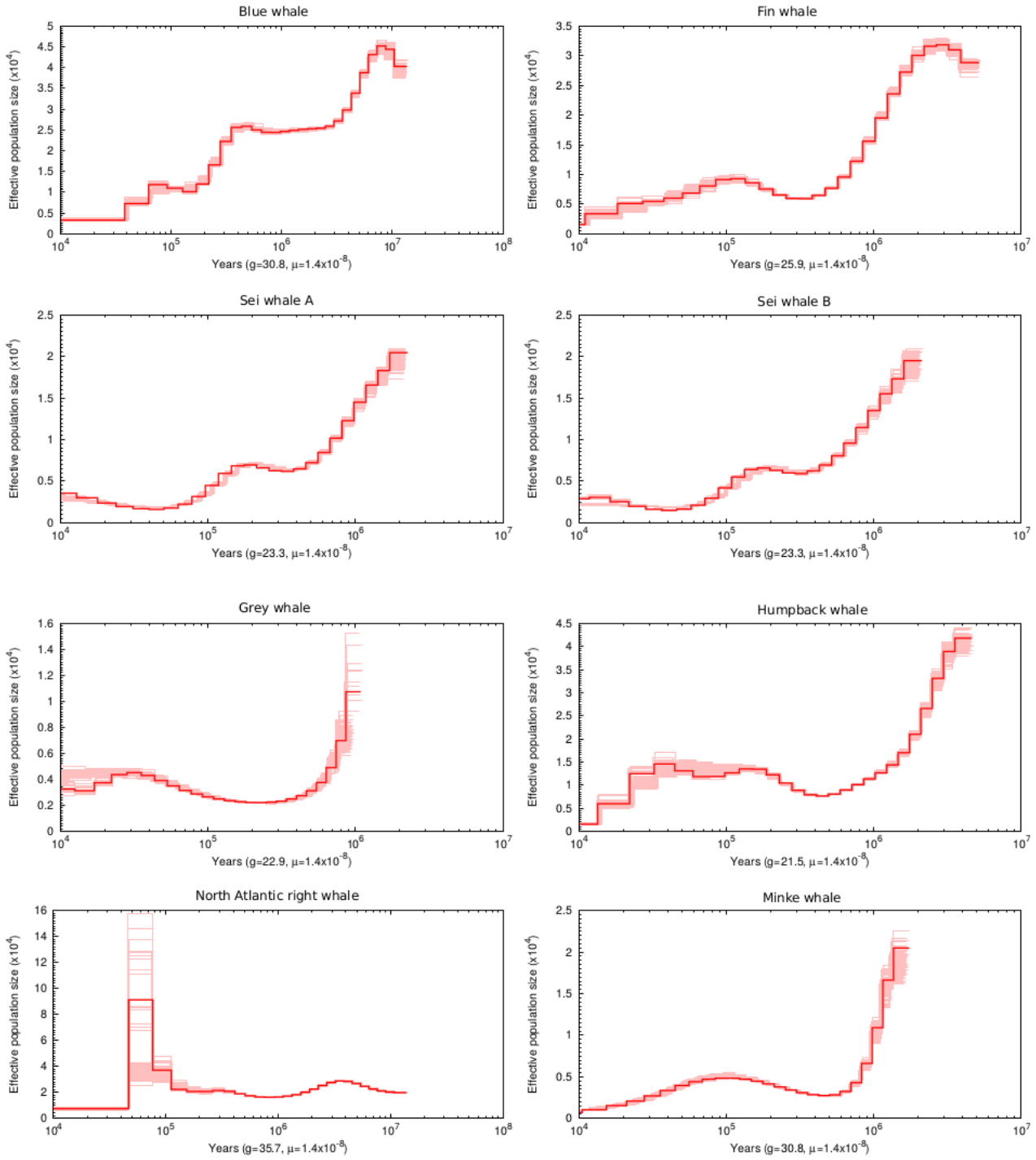

fig. S10. Demographic histories for each individual whale genome with 100 bootstrap replicates. Each panel shows estimated ancestral effective population size calculated by PSMC. Bootstrap replicates are shown in light red. Sample names are given above each panel. Except for North Atlantic right whale, bootstrap replicate indicate little variation in the data. The high degree of variance in bootstrap replicate in North Atlantic right whale likely indicate an artifact in the increase N_e at 10^5 years. g, generation time, μ, mutation rate.

table S1. Sequencing and mapping statistics. For inidividual whale genome sample, a short ID, common species name, number of generated reads, mean insert size and read length as well as statistics after mapping to the bowhead whale genome are shown.

ID	Species	No of reads	% mapped	No. duplicates	Coverage	Insert size	Read-length
Egl00	NA right whale	323,959,050	92.3%	43,364,288	8.32x	470	125
Bbo01	Sei whale	316,038,498	91.4%	55,729,034	7.45x	482	125
Bbo02	Sei whale	316,228,639	91.3%	69,502,856	6.93x	482	125
Bmu00	Blue whale	1,131,873,174	92.9%	106,828,877	27.2x	293	100
Mno00	Humpback whale	549,015,152	91.2%	81,954,390	17.0x	470	125
Bph03	Fin whale	415,640,907	93.2%	58,501,974	10.7x	294	100
Bac00*	Minke whale	248,445,825	91.9%	63,832,762	7.19x	472	100
Ero01	Gray whale	277,126,828	93.5%	52,844,673	6.30x	330	150
Ero02	Gray whale	349,764,342	93.5%	136,428,269	13.1x	323	150

* This library was obtained from Accession No. SRR896642

table S2. Occurrences of repetitive elements in the bowhead whale genome. Repeats are grouped by their respective repeat family. The table shows the number (count), combined lengths und percentage of the genome sequence.

Family	Count	Length (bp)	Genome-%
SINEs	833,543	140,689,989	6.08
LINEs	1,069,602	426,471,811	18.4
LTR	381,792	139,382,337	6.02
DNA	388,951	83,214,371	3.59
Unclassified	7,269	1,327,012	0.06
Small RNA	293,226	65,978,280	2.85
Satellites	118,187	51,524,590	2.23
Simple repeats	503,891	20,650,898	0.89
Low complexity	84,108	4,216,662	0.18

table S3. Number of called substitutions for each whale genome. Fixed sites are the number of nucleotide differences compared to the bowhead whale genome sequence.

Sample	Het. sites	Fixed sites
NA right whale	2,675,248	8,538,671
Sei whale A	1,374,610	27,180,101
Sei whale B	1,256,248	26,394,531
Minke whale	1,200,537	26,446,186
Blue whale	4,371,463	27,158,938
Fin whale	2,668,509	27,549,771
Humpback	2,689,719	28,440,221
Gray whale A	1,591,221	24,091,848
Gray whale B	1,423,890	28,002,269

table S4. Library and sequencing information for the hippopotamus genome assembly. The estimate of the coverage is based on an assumed mammalian genome size of 3.1Gb.

Library	Insert size	No. Reads	No. reads used	Coverage
PE250	250 bp	771,298,964	619,136,098	19.8
PE500	500 bp	479,504,382	392,524,628	12.5
PE800	800 bp	384,369,550	305,871,802	9.77
MP2K	2 kb	343,921,854	28,532,3315	9.11
MP10K	10 kb	102,833,932	81,590,442	2.61

table S5. Summary of repetitive elements in the hippopotamus genome. Repeats are grouped by their respective repeat family. Subfamily counts are given as subsets of their respective families. The table shows the number (count), combined lengths und percentage of the genome sequence. Here, repeat types are also given for subfamilies because the repeats of the hippotamus have not been characterized before.

Repeat type	# elements	lengths (in bp)	% sequence
SINE	877,948	140,668,128	5.80
Alu/B1	27	1,733	0.00
MIRs	530,250	72,759,825	3.00
LINEs	1,168,421	437,621,503	18.0
LINE1	744,824	332,612,025	13.7
LINE2	364,417	91,575,927	3.78
L3/CR1	46,665	10,136,404	0.42
RTE	11,571	3,127,613	0.13
LTR elements	431,887	152,603,906	6.29
ERVL	100777	43,336,314	1.79
ERVL-MaLRs	164,722	56,759,117	2.34
ERV class I	98,204	42,275,346	1.74
ERV class II	40,990	3,107,632	0.13
DNA elements	406,868	84,787,323	3.50
hAT-Charlie	221,833	42,833,726	1.77
TcMar-Tigger	72,285	19,831,649	0.82
Unclassified	7,031	1,301,848	0.05
Total interspersed repeats		816,982,708	33.7
Small RNA	33,7530	66,880,296	2.76
Satellites	125,788	52,266,573	2.15
Simple repeat	54,7104	22,624,259	0.93
Low complexity	102,688	5,155,551	0.21
Total masked			37.0

table S6. A majority-rule consensus analysis of 34,192 individual GF ML trees. Only splits occurring more than 1% are shown. Species in order: 1. Gray whale (A) 2. Gray whale (B) 3. Blue whale 4. Sei whale (A) 5. Sei whale (B) 6. Humpback whale 7. Fin whale 8. Bowhead whale 9. North Atlantic right whale 10. Minke whale.

Set	Count	Frequency
Sets included in the consensus tree		
**........	34192	1.000
*******..*	34192	1.000
...**.....	34190	1.000
.....**...	30755	0.899
..***.....	28662	0.838
*******...	19747	0.578
......	10315	0.302
Sets *not* included in the consensus tree		
..*****...	8918	0.261
*****.....	8721	0.255
..******..*	3507	0.103
..***....*	3410	0.100
.....**..*	2737	0.080
**.......*	2204	0.064
.....*	2118	0.062
*****....*	1711	0.050
***.....	1408	0.041
......	1291	0.038
**...*....	991	0.029
...****...	985	0.029
.**...	915	0.027
..*..**...	907	0.027
***..**...	875	0.026
**....*...	848	0.025
..****....	595	0.017
..***.*...	562	0.016
...**....*	428	0.013
******....	416	0.012
*****.*...	410	0.012
..*......*	396	0.012

Note – The table summarizes the results from the consense analysis. The ranking is according to the number of occurrences of splits. Only splits occurring more frequent than 1% are shown. In each vertical column dots (.) and asterisks (*) represents one individual and its split into the respective group (. or *). For example: row one (`**.........`) has species 1 and 2 (both individuals of gray whale) as the most frequent split against all others, row two (`*******..*`) has species 8 and 9 (bowhead whale and NA right whale) splitting from the other with 34,192 occurrences.

table S7. Common names, scientific names, accession numbers, and source database of additional genomes that were included in the divergence time analyses.

Common name	Scientific Name	Accession	Source
Bajii	*Lipotes vexillifer*	GCF_000442215.1	RefSeq
Bottlenose dolphin	*Tursiops truncatus*	GCA_000151865.3	GenBank
Camel	*Camelus ferus*	GCF_000311805.1	RefSeq
Cow	*Bos taurus*	GCA_000003055.3	GenBank
Dog	*Canis lupus familiaris*	GCA_000002285.2	GenBank
Killer whale	*Orcinus orca*	GCF_000331955.2	RefSeq
Pig	*Sus scrofa*	GCA_000003025.4	GenBank
Sheep	*Ovis aries*	GCF_000298735.1	RefSeq
Sperm whale	*Physeter macrocephalus*	GCA_000472045.1	ENSEMBLE(pre)

table S8. Calibration points used for the divergence time tree, node age estimates in million years ago, and references.

Node	Node age	Reference
Camelidae	63 - 73 Ma	ref. 60
Ruminantia	55 - 60 Ma	ref. 57
Hippopotamus	53.5 - 55 Ma	ref. 29
Cetacea	30.5 - 32.3 Ma	ref. 58
Mysticeti	<28 Ma	ref. 59

table S9. Divergence time estimates for Artiodactyla and Cetacea for nodes in the divergence time tree (Fig. 5). The table shows the mean divergence times, 95% equal-tail confidence interval and 95% highest posterior density. For comparison divergence time estimates from ref. 8 and ref. 21 are given if present in the respective studies. Times are given in million years ago (Ma).

Node	Description	Mean	95% Equal-tail	95 % HPD	Arnason (2004)	McGowen (2009)
t_n25	Hippopotamidae	54.2	0.535 - 0.550	0.535 - 0.550	53.30	-
t_n26	Cetacea	31.8	0.307 - 0.324	0.308 - 0.324	35.00	36.36
t_n27	Odontoceti	27.6	0.212 - 0.310	0.226 - 0.315	32.1 ± 1.7	34.69
t_n28	Delphinidae + Lipotiidae	18.7	0.099 - 0.262	0.101 - 0.264	22.4 ± 1.5	24.70
t_n29	Delphinidae	8.81	0.035 - 0.165	0.030 - 0.158	-	10.80
t_n30	Mysticeti	28.3	0.275 - 0.295	0.274 - 0.294	20.90	28.79
t_n31	Balaenopteroidae	10.5	0.053 - 0.204	0.045 - 0.189	12.70	13.80
t_n32	Balaenopteridae ex. B. borealis	8.35	0.042 - 0.162	0.035 - 0.147	-	10.21
t_n33	Eschrichtiidae + Megaptera + B. physalus	7.49	0.036 - 0.148	0.031 - 0.135	-	9.04
t_n34	Megaptera + B. physalus	4.98	0.018 - 0.108	0.014 - 0.097	-	7.06
t_n35	Eschrichtius	0.08	0.000 - 0.002	0.000 - 0.002	-	-
t_n36	B. borealis + B. musculus	5.79	0.022 - 0.125	0.016 - 0.111	10.30	8.74
t_n37	B. borealis	0.29	0.001 - 0.008	0.001 - 0.007	-	-
t_n38	Balaenidae	4.38	0.012 - 0.126	0.007 - 0.103	-	5.38