Deciphering the evolutionary signatures of pinnipeds using novel genome sequences: The first genomes of *Phoca largha*, *Callorhinus ursinus*, and *Eumetopias jubatus*

Jung Youn Park¹, Kwondo Kim²,³, Hawsun Sohn⁴, Hyun Woo Kim⁴, Yong-Rock An⁵, Jung-Ha Kang⁴, Eun-Mi Kim², Woori Kwak², Chul Lee², DongAhn Yoo²,³, Jaehoon Jung³,⁶, Samsun Sung³, Joong Yoon² & Heebal Kim²,³,⁶

The pinnipeds, which comprise seals, sea lions, and walruses, are a remarkable group of marine animals with unique adaptations to semi-aquatic life. However, their genomes are poorly characterized. In this study, we sequenced and characterized the genomes of three pinnipeds (*Phoca largha*, *Callorhinus ursinus*, and *Eumetopias jubatus*), focusing on site-wise sequence changes. We detected rapidly evolving genes in pinniped lineages and substitutions unique to pinnipeds associated with amphibious sound perception. Phenotypic convergence-related sequence convergences are not common in marine mammals. For example, *FASN*, *KCNA5*, and *IL17RA* contain substitutions specific to pinnipeds, yet are potential candidates of phenotypic convergence (blubber, response to hypoxia, and immunity to pathogens) in all marine mammals. The outcomes of this study will provide insight into targets for future studies of convergent evolution or gene function.

Marine mammals are a classic example of convergent evolution in terms of adaptation of terrestrial mammals to the marine environment. During secondary adaptation to the marine environment, marine mammals experienced similar environmental challenges, which have resulted in shared morphological or physiological features across distant taxa. For instance, they have experienced similar changes in skin and limbs, and subsequently became streamlined¹². Adaptive traits related to hypoxia are shared features of marine mammals¹³. Marine mammals include three orders: cetaceans (whales, dolphins, and porpoises), pinnipeds (seals, sea lions, and walruses), and sirenians (manatees and dugongs)⁴. They have evolved to inhabit the ocean in multiple lineages. Cetaceans and sirenians emerged around 40–50 million years ago (mya) from Cetartiodactyla and Afrotheria, respectively⁵. Pinnipeds emerged within the Carnivora approximately 20 million years later⁶. This implies that different molecular changes occurred across separate lineages, possibly resulting in divergent phenotypic changes. However, most studies related to marine mammals have focused on convergent evolution, although some of the adaptations of marine mammals to an aquatic lifestyle vary among species⁷.

¹Biotechnology Research Division, National Institute of Fisheries Science, 216 Haean-ro, Gijang-eup, Gijang gun, Busan, 46083, Republic of Korea. ²Interdisciplinary Program in Bioinformatics, Seoul National University, Kwan-ak Gu, Seoul, Republic of Korea. ³C&K genomics, C-1008, H businesspark, 26, Beobwon-ro 9-gil, Songpa-gu, Seoul, Republic of Korea. ⁴Cetacean Research Institute, National Institute of Fisheries Science, 250 Jangsaengpo Gorae-ro, Nam-gu, Ulsan, 44780, Republic of Korea. ⁵Department of Taxonomy and Systematics, National Marine Biodiversity Institute of Korea, eocheon-gun, Chungcheongnam-do, 33662, Republic of Korea. ⁶Department of Agricultural Biotechnology, Seoul National University, Kwan-ak Gu, Seoul, Republic of Korea. Jung Youn Park and Kwondo Kim contributed equally. Correspondence and requests for materials should be addressed to H.K. (email: heebal@snu.ac.kr)
Pinnipeds, which consist of three families (Phocidae, Otariidae, and Odobenidae) are distinguishable from other marine mammals. Most pinnipeds are semi-aquatic, unlike other marine mammals that spend their entire lives in the water, and have modified limbs as flippers that propel them both in the water and on land. In addition, with the exception of the walrus, which is the only extant species of the family Odobenidae, all pinnipeds have fur coats. These distinct characteristics have not been sufficiently characterized at the molecular level. Although a draft fur seal genome has recently been assembled, the evolutionary and biological aspects of pinnipeds have not been investigated. Indeed, the genome of the Weddell seal (family Phocidae) has not been completed. In addition, most phylogenetic studies of pinnipeds have used limited marker sequences, such as that of the mitochondrial genome.

Comparative genomics enables investigation of the convergent evolution of distant species. For example, convergent amino acid changes for vocal learning were identified by sequencing 48 avian genomes. Similarly, Parker et al. reported nearly 200 convergent loci in the genomes of echolocating mammals. Although there are more studies to demonstrate to phenotypic convergence-linked sequence convergence, molecular convergence toward phenotypic convergence, at least in marine mammals, seems to be uncommon. By analyzing 22 mammalian genomes, including those of three marine mammals, Foote et al. suggested that different molecular pathways could be used to reach the same phenotype. In a study of the Hox gene family in mammals, only a fraction of sites had positive selection signatures shared by three independent marine mammal lineages. Rather than sequence-level, gene-level convergence was presented as widespread signatures when evolution was used. Therefore, there is convergence at the functional level or higher in separate mammalian lineages, and different marine mammal lineages have used different molecular pathways to achieve phenotypic convergence.

Here, we constructed draft genomes of three species of two pinniped families: Phoca largha (Phocidae) and Callorhinus ursinus (Odobenidae) (Fig. S1 and Supplemental Note S1). We identified genes with a positive selection signature that were common to the three pinnipeds but absent from other mammals, which are likely related to the unique traits of pinnipeds. In addition, divergent molecular changes likely to occur only in the pinniped lineage during phenotypic convergence of marine mammals were investigated.

Results

Genome assembly and annotation. Before assembling the genomes of the three pinnipeds, we estimated the genome sizes using the 19-mer distribution of paired-end reads. The estimated genome sizes were 2.61, 2.71, and 2.64 Gbp for the spotted seal (SS), northern fur seal (NFS), and Steller sea lion (SSL), respectively (Fig. S2). The genomic DNA of the three pinnipeds was assembled to a size of approximately 2.5 Gbp, which is similar to that of previously assembled genomes (Antarctic fur seal, Hawaiian monk seal, and Weddell seal). Summary statistics of the final assembly are provided in Table S1. To assess the quality of the draft genomes, we remapped paired-end reads with a 350 bp insert size, which yielded alignment rates of >98% for the three genomes (98.24, 98.74, and 98.73% for SS, NFS, and SSL, respectively). The completeness of core-orthologs was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO). Each of the three genomes contained more than 90% core-orthologs from the class Mammalia, in the form of either complete or fragmented sequences (Table S2). The GC contents of the three genomes were investigated using 500 bp bins, and were similar to those of the draft genomes of related species (Fig. S3).

Repeat elements accounted for 35.83, 40.40, and 35.78% of the SS, NFS, and SSL genomes, respectively. Of the repeat regions, long interspersed nuclear element (LINE) was the most extended element in terms of base pairs (Table S3). After mapping the identified repeat elements, 33,988, 32,740, and 28,081 protein-coding genes were predicted for SS, NFS, and SSL, respectively (Table S4). Of the predicted genes, ~95% were functionally annotated to at least one of the InterPro, SwissProt, and TrEMBL databases (Table S5).

Therefore, the SS, NFS, and SSL genomes were not significantly different from one another in terms of various statistics related to genome assembly. Because the three species are related, this similarity suggests that the three genomes have similar levels of completeness.

Phylogenomics and protein-coding gene families. To identify the relationships among SS, NFS, and SSL and other related species, we constructed a maximum-likelihood (ML) tree using the amino acid sequence of one-to-one orthologs generated using a dataset of the proteomes of nine species available in public databases. In total, there were 2,907 one-to-one orthologs, the combined length of which was 982,250 amino acid residues. The newly constructed tree provided robust support for the known phylogenetic tree of marine mammals (Fig. 1A) and the phylogenetic tree is used in the downstream analysis for positively selected genes and substitutions.

We constructed orthologous gene clusters using the genomes of six marine mammals to identify gene clusters and their functions unique to pinnipeds (Fig. S4). The pinniped genomes contained 13,919 (NFS), 13,441 (SS), and 14,165 (SSL) orthologous gene families, respectively, 9,639 of which were shared by all three pinnipeds (Fig. 1B). Of these gene families, 1,874 were present in all pinnipeds, but not in three other mammals. By Gene Ontology (GO) enrichment analysis, we found these gene families to be enriched in 31 terms (p-value < 0.05), several of which were related to an aquatic lifestyle, such as ‘aorta development’, ‘sterol biosynthetic process’, ‘cardiac septum development’, ‘coronary vasculature development’, and ‘cellular response to oxidative stress’ (Table S6).

To investigate gene-family expansion and contraction, a computational analysis of gene-family sizes using the orthologous gene clusters was performed in CAFE. By comparing six marine mammals, we found that 874 gene families were expanded, while 1,925 gene families were contracted in the pinniped lineage. Of these gene families, a subset of the Protocadherin (Pcdh) family (herein named family 34) was significantly expanded in the pinniped lineage (p = 0.000346). The genomes of the pinnipeds contained a larger number of Pcdh genes than those of the
other marine mammals (Fig. 1C). Pcdhs are the largest mammalian subgroup of the cadherin superfamily, and have functions associated with the nervous system such as in olfactory sensory neurons. The number of Pcdhs varies among vertebrate lineages.

Genes with accelerated evolution in the pinniped lineage. To detect positive selection in the pinniped lineage, a dN/dS analysis using the branch-site model was performed. The branch-site model allows dN/dS (ω) to vary both among sites in the protein and across branches on the tree. Therefore, we hypothesized a few sites in the pinniped branches to have different ω ratios compared to other branches and that the genes containing these sites might be related to the unique features of pinnipeds. After the filtering step (see Methods), we analyzed 2,754 one-to-one orthologs identified in the proteomes of 12 mammals, of which seven genes with 145 sites were under positive selection (Bonferroni-corrected p < 0.05, posterior probability based on Bayes empirical Bayes inference [BEB] > 0.95; Table 1). Of these genes, transmembrane protein 132B (TMEM132B) contained the largest number of positively selected sites (52 sites). Of the seven genes, six contained 29 conserved domains with 74 sites (51%) under positive selection. GO terms were assigned to each gene, and the following functional associations with pinniped lifestyle were found: TECTA, sensory perception of sound (GO:0041593), SPEG, muscle organ development (GO:0007517), and ADAMTS5, defence response to bacterium (GO:0042742) and tooth eruption (GO:0044691). TECTA encodes alpha-tectorin, a major non-collagenous glycoprotein of the tectorial membrane, an extracellular matrix in the inner ear. Mutations in TECTA result in hearing loss (OMIM: 602574). SPEG is required for cardiac development and is associated with cardiac myopathy (OMIM: 615950). ADAMTS5, which encodes an extracellular matrix-degrading enzyme, plays an important role in the T-cell immune response to viral infection.

To assess their uniqueness, the amino acid residues positively selected in the pinniped lineage were compared to other species in our analysis as well as in publicly available databases. For example, we investigated 4 of the 18 sites within TECTA after manually filtering out amino acid residues with spurious alignment (Fig. 2A). The four sites were pinniped-specific compared to the other nine species (Fig. 2B). Moreover, a 100-way multi-alignment showed that two pinnipeds (Pacific walrus and Weddell seal) had residues identical to those in the three pinnipeds

Figure 1. Phylogenomics and protein-coding gene families of pinnipeds. (A) Species tree of 12 terrestrial and marine mammals constructed by the maximum-likelihood method. (B) Orthologous gene clusters in three pinnipeds. (C) Number of intact (coverage ≥ 90%) and partial (coverage < 90%) genes that belong to Protocadherin gene families, named family 34 in our dataset (Dol, dolphin; Man, manatee; Min, Minke whale; Nor, northern fur seal; Spo, spotted seal; Ste, Steller sea lion).
in this study (Fig. S5). We could only find a small number of residues matching those in 100 vertebrates at these sites (Fig. S5). Consequently, the four sites within TECTA might be unique to pinnipeds and generated during their adaptation to a semi-aquatic environment.

Unique substitutions of pinnipeds contributed to the phenotypic convergence of marine mammals. Parallel substitutions are widespread in marine mammals; however, most are not unique to marine mammals. In this study, about half of the parallel substitutions shared by marine mammals were also found in terrestrial mammals, and a considerable number of unique substitutions was found between species with no obvious phenotypic convergence (Figs S6–S8). Therefore, we hypothesized the existence of pinniped-specific substitutions that contributed to aquatic adaptation and are shared by marine mammals.

First, we focused on gene-level convergence (Fig. S9) and conducted a dN/dS analysis of one-to-one orthologs using the branch model. The branch model allows the dN/dS (ω) ratio to vary among branches in the phylogeny and is useful for detecting positive selection acting on particular lineages. In this way we aimed to detect candidate genes with different ω ratios among the marine mammal lineages rather than candidate sites, which may contribute to phenotypic convergence among marine mammals. Of the 2,754 filtered one-to-one orthologs, the branch model-based dN/dS analysis detected 853 positively selected genes in marine mammal lineages (Fig. S10b, cetaceans, pinnipeds, and sirenians, Bonferroni corrected p-value < 0.05). These are hereafter referred to as rapidly evolving genes (REGs). A subset of 853 REGs covered the following functional categories possibly associated with marine mammals’ adaptation to the ocean: muscle physiology (GO:0007015, GO:0035914, GO:0007507), lipid metabolism (GO:0006629, GO:0006869, GO:0016042), sensory system (GO:0007605, GO:0042472, and GO:0021772), skin and connective tissue (GO:0008544, GO:0043588, and GO:0030216), cardiovascular system (GO:0086091, GO:0060976, and GO:0007507), and resistance to oxidative stress (GO:0001666).

Table 1. Genes with accelerated evolution in the pinniped lineage. H1_fg_omega: dN/dS value (ω) on foreground given H1 (ω varies across the branches); H0_lnl: log likelihood given H0 (ω does not vary across the branches); H1_lnl: log likelihood given H1; H0_lnl: log likelihood given H0. *Number of positively selected sites with a BEB of >0.95. The numbers of positively selected sites within domain regions are shown in parentheses.

| Gene     | H1_fg_omega (ω) | Proportion (H1) (1 – p0 = p1/p0) | H0_lnl | H1_lnl | Likelihood ratio | p-value | Adjusted p-value | # of positively selected sites* |
|----------|----------------|---------------------------------|--------|--------|-----------------|---------|-----------------|-------------------------------|
| TMEM132B | 3.81581        | 0.01666                         | −6438.78 | −6419.68 | 38.20475        | 6.37E-10 | 1.18E-06        | 52 (22)                       |
| PARP1    | 4.76894        | 0.00604                         | −5357.53 | −5341.29 | 32.48145        | 1.26E-08 | 2.22E-05        | 23 (22)                       |
| TECTA    | 3.67139        | 0.00194                         | −12076.1 | −12060.4 | 31.42787        | 2.07E-08 | 3.83E-05        | 18 (14)                       |
| FUBP3    | 4.89809        | 0.01916                         | −4880.95 | −4869.76 | 22.38143        | 2.24E-06 | 0.004144        | 12 (1)                        |
| IGFBP1   | 4.96893        | 0.00201                         | −4448.2 | −4438.13 | 20.13898        | 7.02E-06 | 0.01332         | 19 (2)                        |
| NPEG     | 4.81594        | 0.00254                         | −11218.8 | −11209.4 | 18.85029        | 1.41E-05 | 0.026085        | 13 (13)                       |
| ADAMTSS  | 4.38148        | 0.00124                         | −4320.48 | −4311.43 | 18.1014         | 2.09E-05 | 0.038665        | 8 (0)                         |

Table 1. Genes with accelerated evolution in the pinniped lineage. H1_fg_omega: dN/dS value (ω) on foreground given H1 (ω varies across the branches); H0_lnl: log likelihood given H0 (ω does not vary across the branches); H1_lnl: log likelihood given H1; H0_lnl: log likelihood given H0. *Number of positively selected sites with a BEB of >0.95. The numbers of positively selected sites within domain regions are shown in parentheses.

Figure 2. Results of a branch-site model analysis of TECTA. (A) Bayes empirical Bayes (BEB) posterior probability in TECTA. Shaded area, conserved domain regions. (B) Sequence of sites with significant BEB (>0.95). Red and blue shaded areas, pinnipeds and other mammals, respectively.

...
We also calculated the site-wise log likelihood support (SSLS) values for the amino acid sequences of 2,754 genes (996,522 residues in total) and calculated the ΔSSLS values to detect site-wise signatures of divergent evolution. The ΔSSLS value is indicative of the goodness-of-fit of each site to a pair of phylogenetic trees. We aimed to detect genes positively selected in three marine mammal lineages with substitutions unique to pinnipeds. Therefore, we calculated the SSLS for two hypotheses: H₀, divergence among marine mammal clades and H₁, convergence among marine mammal clades. Therefore, a ΔSSLS (log likelihood of H₀ − log likelihood of H₁) value > 0 means that the site in question supports divergence among marine mammal clades. We used the ΔSSLS value as a filtering criterion to exclude sites supporting convergence among marine mammals. By excluding those with low ΔSSLS values, we identified pinniped-specific sites that support the separation clades of marine mammals. We expected that this analysis would generate more reliable sites than directly extracting unique substitutions over REGs, as it considers the overall phylogeny not just the sequence itself.

We regarded the 9,965 residues with the top 1% ΔSSLS values as being supported by divergent substitutions (support for H₀) rather than convergent substitutions among three marine mammal clades (support for H₁) (Fig. 3A). We termed the 2,159 genes containing at least one of these residues as divergent substitution genes (DSGs). DSGs covered most of the 2,754 one-to-one orthologs (78%), and 85% of total residues had positive ΔSSLS values. Therefore, the majority of the sequences supported the commonly accepted phylogeny.

Of the 853 REGs, 658 (3,277 residues) had at least one top 1% ΔSSLS site (Fig. 3B). Although these genes covered the functional categories associated with marine mammals’ adaptation, a single residue divergence supported by likelihood divergence (ΔSSLS) could be vulnerable to sequencing error. We also focused on sequence changes common to the pinniped clade; that is, changes from the ancestral node sequence shared by dog to that of the node of each pinniped. Therefore, we investigated unique substitutions (any amino acid residue at the same position in all three pinnipeds that was found in neither the ancestral nodes with their respective terrestrial taxa [dog] nor in other mammals) to rule out sequence divergences other than ancestral substitutions unique to the pinniped clade. There were 1,731 genes with at least one unique substitution (7,878 residues); these were termed unique substitution genes (USGs), 63 of which contained top 1% ΔSSLS residues at the same positions as unique substitutions. Finally, we obtained 24 REGs containing top 1% ΔSSLS residues and unique substitutions at the same positions (Fig. 3 and Table 2).

Although the 24 REGs are supported by rapid evolutionary rates (dN/dS) and fixation of amino acid residues within the pinniped clade, the precise phenotypic effects of the unique substitutions cannot currently be ascertained. However, several of the 24 REGs have known functional associations that suggest a role in the convergent phenotypic evolution of marine mammal lineages. For instance, FASN encodes fatty-acid synthase, which catalyzes the conversion of acetyl-CoA and malonyl-CoA to long-chain saturated fatty acids and is related to obesity. KCNA5 (potassium voltage-gated channel subfamily A member 5) encodes voltage-gated potassium...
particular frequencies. Interestingly, the four positively selected sites in TECTA are responsible for loss of hearing at the TECTA of pinnipeds by tuning their hearing ranges. Indeed, mutations in TECTA challenged by the need to function efficiently underwater and in air, unlike the solely underwater hearing of vertebrates (Fig. S5). Although its relationship with amphibious sound perception is unclear, TECTA found in the pinniped lineage should be conserved. Comparative analysis of amphibious sound perception is unclear, TECTA should be investigated in future studies of amphibious sound perception in pinnipeds. The pinniped lifestyle might influence the function of other candidate genes, such as SPEG and ADAMTS5. Comparative analysis of amphibious mammals may reveal their adaptations at the molecular level and show that an amphibious lifestyle results in selection pressure. We found that a considerable number of parallel substitutions are not unique to marine mammals, consistent with two recent reports. This implies that molecular convergence is not a driving force of phenotypic convergence among marine mammals, and that different clades of marine mammals used different molecular pathways to reach similar phenotypes. Although this phenomenon has been observed several times in marine mammals, whether it also applies to other clades is unclear. More evidence in other clades is needed to generalize this phenomenon to other forms of phenotypic convergence. Because sequence convergences leading to phenotypic convergences are not common, we assumed that unique substitutions contributed to the aquatic adaptation of pinnipeds. In our analyses, three genes, FASN,
**Ortholog identification.** The complete proteome datasets were downloaded from UCSC Genome Browser for the following nine mammals: human (hg19), mouse (mm10), dog (canFam3), cow (bosTau8), manatee (triMan1), dolphin (turTru2), Minke whale (balAcu1), opossum (monDom5), and elephant (loxAfr3). Gene clusters for these nine mammals and three pinnipeds were identified using OrthoMCL with the -q, no_is options.

For the three gene sets, the best match of a BLASTP search against the SwissProt and TrEMBL databases was assigned to putative functions. Gene motifs and domains were determined using InterProScan v. 5.19. The GO IDs for each gene were obtained from the corresponding InterPro entries.

**Methods**

**Ethics statement.** No ethics approval was required for the collection of DNA from blood samples of bycaught carcasses.

**Sample information and collection.** We collected five pinniped samples from Korean waters. Three male Northern fur seals (Callorhinus ursinus) were bycaught in set nets and collected during January and February 2016 (one was used to produce sequence data). A bycaught female Steller sea lion (Eumetopias jubatus) was collected in April 2008. A female spotted seal (Phoca largha) was collected on a beach in August 2015. All of the above were found in the waters off Gangwon-do, northeastern South Korea.

**DNA sequencing and genome assembly.** For whole-genome shotgun sequencing and draft genome assembly, we constructed two paired-end libraries with insert sizes of 350 and 700 bp using the Illumina TruSeq DNA Sample Preparation Kit (Illumina, San Diego, CA, USA). For the Steller sea lion genome, mate-pair libraries with insert sizes of 3, 9, and 40 kb were constructed as scaffolds using the Illumina Nextera mate-pair library construction protocol (Illumina). Sequence reads were generated using the Illumina Nextseq 500 platform. Information on the constructed libraries and sequencing data is provided in Table S7.

The 19-mer distribution of the paired-end library with an insert size of 350 bp was calculated using Jellyfish, and the sizes of three genomes were estimated (Fig. S1). To retrieve high-quality sequence reads, the quality of the raw data was controlled using FASTQC. Artifact sequences were removed via Trimmomatic for paired-end libraries, and Nxtrim for mate-pair libraries. Sequencing errors within each read were estimated and discarded using the error-correction module of Allpaths-LG. We assembled error-corrected paired-end reads using IDBA_UD with the option of pre-correction and kmin = 40. Scaffolding on initial contigs was conducted using the paired-end reads with a 700 bp insert size, and mate-pair reads sequentially by SSPACE and ScaffMatch. After scaffolding, we iteratively filled gaps using Gapcloser with the -l 155 and -p 31 parameters.

RepeatModeler, which includes RECON, RepeatScout, and TRF, was used to create a custom database for each species. A custom library was constructed by integrating the custom databases into the Repbase database of mammals. Repeat elements were identified and masked using RepeatMasker with the -c, no_is options.

**Genome annotation.** Two approaches were used to predict protein-coding genes. First, manually curated protein sequences of Mammalia were retrieved from Swiss-Prot and aligned to the pinniped genomes using tblASTN. The homologous genome sequences with E-values ≤ 1E-5 were extracted and realigned to the matched proteins using Exonerate to predict splice sites. Ab initio gene prediction was conducted using Augustus, Geneid, and GlimmerHMM software with the default options. Predicted genes using each approach were combined using EvidenceModeler into a consensus gene set.

For assessment of the quality of the draft genome, we remapped paired-end reads with a 350 bp insert size and investigated completeness of core-orthologs using BUSCO. For the three gene sets, the best match of a BLASTP search against the SwissProt and TrEMBL databases was assigned to putative functions. Gene motifs and domains were determined using InterProScan v. 5.19. The GO IDs for each gene were obtained from the corresponding InterPro entries.
Phylogenomic analyses using a genome-wide set of one-to-one orthologs. Amino acid sequences of 12 mammals corresponding to the one-to-one orthologs were individually aligned using ClustalW v. 2.172. A concatenated alignment was then prepared by merging individual alignments. The concatenated alignment was trimmed using Gblocks v. 0.91b73 with auto settings.

The best-fit substitution model for the alignment was determined using ModelGenerator29. For phylogenetic analyses, RAXML v. 7.2.880 was used to generate ML trees. Rapid bootstrap analysis and identification of the best-scoring ML tree (-f a option) were performed using RAXML v. 7.2.880. Bootstrap support values/percentages were determined using 100 replicates. A Jones-Taylor-Thornton amino acid substitution model81 (with the PROCATIITTF option) as recommended by ModelGenerator82 was used to construct the ML trees.

Detection of lineage-specific gene losses and gains. Using the gene clusters defined by Orthomcl v. 2.0.976, the genes in each gene family group were enumerated and converted to input data for CAFÉ software v. 3.117. Expansion or contraction of the gene families was defined by comparing the cluster size of the ancestor to that of each of the current species using CAFÉ17.

Detection of positively selected genes and substitutions. To detect positively selected genes, coding sequence alignments were prepared by pal2nal v. 1483 using the amino acid alignments of the one-to-one orthologs. After trimming of the poorly aligned regions, alignments that are shorter than 100 bp or contain an internal stop codon were excluded.

To detect positive selection affecting a few sites in particular lineages (foreground branches, pinniped lineage in this study), we employed a branch-site model, which allows the ω ratio to vary both among lineages and among sites. We used the ML method of codeml in PAML v. 4.984, which estimates the rate of non-synonymous substitutions (dN), the rate of synonymous substitutions (dS), and the ratio of the non-synonymous to synonymous substitution rates (ω) values using the F3X4 codon frequencies. An alternative codon substitution model was specified using model = 2, dNsites = 2 (model A

Calculation of site-wise likelihood support. To detect sites with molecular divergence that supported the monophyly of pinnipeds, we fitted the amino acid sequence alignment of one-to-one orthologs to a null model (H0, species tree) and an alternative model (H1, monophyly of marine mammals) (Fig. 3A). The goodness-of-fit of each site to a pair of phylogenetic trees under a given model was calculated as the SSLS value and directly compared as ΔSSLS = lnL(H0) - lnL(H1). Positive ΔSSLS values indicate a better fit of the model to the species tree, H0 (supporting divergence), whereas negative ΔSSLS values indicate a better fit to H1 (supporting convergence). The substitution model for each gene was determined by ModelGenerator85. The SSLS value for each site of alignment was estimated by RAxML v. 7.2.880.

Identification of parallel and unique substitutions. We defined parallel substitutions as any amino acid change at the same position in marine mammals different from that of the ancestral node of each marine group, but identical in the three marine groups. To identify parallel amino acid changes in marine mammals, the species tree constructed in this study was used to reconstruct the ancestral sequences. The ancestral sequences for each node were reconstructed by Joint method using FastML v. 3.186. We allowed FastML 3.1 to estimate the branch length of the phylogenetic tree for each gene when the ancestral sequences were reconstructed using the set of 12 mammals. For the sites with parallel and unique substitutions, the amino acid sequences of 100 vertebrates were investigated by 100-way multi-alignment87 with the UCSC genome browser.

Conserved domain search. To determine whether positively selected sites are located in gene functional domains, we searched for conserved domains within positively selected genes using the CD-Search tool in the NCBI88. The amino acid sequences of human orthologs were used as a query set with the following settings: data source, CDD v. 3.16; expected value threshold, 0.01; composition-based statistical adjustment, applied; low-complexity filter, not applied.

Gene ontology analysis. We mapped the identified genes to GO categories in Ensembl89 to identify those putatively associated with a specific function, such as adipose tissue development. Gene set enrichment tests were performed by DAVID functional annotation90 using a cutoff P-value of <0.05.

Availability of Data and Material
The datasets generated during the current study are available in the NCBI repository, PRJNA422019.
References

1. Fish, E. E., Howe, L. E. & Murray, M. M. Hydrodynamic flow control in marine mammals. *Integrative and Comparative Biology* **48**, 788–800 (2008).

2. Chikina, M., Robinson, J. D. & Clark, N. L. Hundreds of genes experienced convergent shifts in selective pressure in marine mammals. *Molecular biology and evolution* **33**, 2182–2192 (2016).

3. Andersen, H. T. Physiological adaptations in diving vertebrates. *Physiological Reviews* **46**, 212–243 (1966).

4. Jefferson, T. A., Leatherwood, S. & Webber, M. A. Marine mammals of the world. (Food & Agriculture Org. 1993).

5. Berta, A., Sumich, J. L. & Kovacs, K. M. Marine mammals: evolutionary biology. (Academic Press 2005).

6. Berta, A. Pinnipedia, overview. *J. Zool.* **203**, 1525–1531 (2002).

7. Rybchynsky, N., Dawson, M. R. & Tedrow, R. H. A semi-aquatic Arctic mammalian carnivore from the Miocene epoch and origin of Pinnipedia. *Nature* **458**, 1021 (2009).

8. Riedman, M. The pinnipeds: seals, sea lions, and walruses. (Univ of California Press 1990).

9. Humble, E. et al. A draft fur seal genome provides insights into factors affecting SNP validation and how to mitigate them. *Molecular ecology resources* **16**, 999–921 (2016).

10. Slade, R. W., Moritz, C. & Heideman, A. Multiple nuclear-gene phylogenies: application to pinnipeds and comparison with a mitochondrial DNA gene phylogeny. *Molecular Biology and Evolution* **11**, 341–356 (1994).

11. Davis, C. S., Delisle, J., Stirling, I., Siniff, D. B. & Strobeck, C. A phylogeny of the extant Phocidae inferred from complete mitochondrial DNA coding regions. *Molecular phylogenetics and evolution* **33**, 363–377 (2004).

12. Fulton, T. L. & Strobeck, C. Multiple markers and multiple individuals refine true seal phylogeny and bring molecules and morphology back in line. *Proceedings of the Royal Society of London B: Biological Sciences* **277**, 1065–1070 (2010).

13. Zhang, G. et al. Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).

14. Parker, J. et al. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502** (2013).

15. Foote, A. D. et al. Convergent evolution of the genomes of marine mammals. *Nature genetics* **47**, 272–275 (2015).

16. Nery, M. F., Borges, B., Dragalzew, A. C. & Kohlsdorf, T. Selection on different genes with equivalent functions: the convergence story told by Hox genes along the evolution of aquatic mammalian lineages. *BMC evolutionary biology* **16**, 113 (2016).

17. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269–1277 (2006).

18. Hulipiau, P. & Van Roy, F. Molecular evolution of the cadherin superfamily. *The international journal of biochemistry & cellular biology* **41**, 349–369 (2009).

19. Wang, X. et al. Gamma protocadherins are required for survival of spinal interneurons. *Neuron* **36**, 843–854 (2002).

20. Chen, W. V. et al. Functional significance of isoform diversification in the protocadherin gamma gene cluster. *Neuron* **75**, 402–409 (2012).

21. Hasegawa, S. et al. The protocadherin-α family is involved in axonal coalescence of olfactory sensory neurons into glomeruli of the olfactory bulb in mouse. *Molecular and Cellular Neuroscience* **38**, 66–79 (2008).

22. Yagi, T. Clustered protocadherin family. Development, growth & differentiation **50** (2008).

23. Yang, Z., Wong, W. S. & Nielsen, R. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular biology and evolution* **22**, 1107–1118 (2005).

24. Collin, R. W. et al. Mid-frequency DFNA8/12 hearing loss caused by a synonymous TECTA mutation that affects an exonic splice enhancer. *European journal of human genetics* **16** (2008).

25. Meyer, N. C. et al. Identification of three novel TECTA mutations in Iranian families with autosomal recessive nonsyndromic hearing impairment at the DFNB21 locus. *American Journal of Medical Genetics Part A* **143**, 1623–1629 (2007).

26. Alasti, F. et al. A novel TECTA mutation confirms the recognizable phenotype among autosomal recessive hearing impairment families. *International journal of pediatric otorhinolaryngology* **72**, 249–255 (2008).

27. Liu, X. et al. Disruption of striated preferentially expressed gene locus leads to dilated cardiomyopathy in mice. *Circulation* **119**, 261–268 (2009).

28. Agrawal, P. B. et al. SPEG interacts with myotubularin, and its deficiency causes centronuclear myopathy with dilated cardiomyopathy. *The American Journal of Human Genetics* **95**, 218–226 (2014).

29. Stambas, J. et al. (Am Assoc Immnol 2017).

30. McMahon, M. K., McCulloch, D. & Stambas, J. (Am Assoc Immnol 2016).

31. Zhou, X., Seim, I. & Gladyshev, V. N. Convergent evolution of marine mammals is associated with distinct substitutions in common genes. *Scientific reports* **5** (2015).

32. Yang, Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular biology and evolution* **15**, 568–573 (1998).

33. Wakil, S. J. Fatty acid synthase, a proficient multifunctional enzyme. *Biochemistry* **28**, 4523–4530 (1989).

34. Lofthus, T. M. et al. Reduced food intake and body weight in mice treated with fatty acid synthase inhibitors. *Science* **288**, 2379–2381 (2000).

35. Firth, A. L. et al. Hypoxia selectively inhibits KCN5A channels in pulmonary artery smooth muscle cells. *Annals of the New York Academy of Sciences* **1177**, 101–111 (2009).

36. Platoshyn, O. et al. Acute hypoxia selectively inhibits KCN5A channels in pulmonary artery smooth muscle cells. *American Journal of Physiology-Cell Physiology* **290**, C907–C916 (2006).

37. Bär, E., Whitney, P. G., Moor, K., e Sousa, C. R. & LeibundGut-Landmann, S. II-17 regulates systemic fungal immunity by controlling the functional competence of NK cells. *Immunity* **40**, 117–127 (2014).

38. Cypowyj, S., Picard, C., Marodi, I., Casanova, J. L. & Puel, A. Immunity to infection in IL-17-deficient mice and humans. *European journal of immunology* **42**, 2246–2254 (2012).

39. Reichmuth, C., Holt, M. M., Mulsow, J., Sills, J. M. & Southall, B. L. Comparative assessment of amphibious hearing in pinnipeds. *Journal of Comparative Physiology A* **199**, 491–507 (2013).

40. Wartozik, D. & Ketten, D. R. Marine mammal sensory systems. *Biology of marine mammals* **1**, 117 (1999).

41. Verhoeven, K. et al. Mutations in the human α-tectorin gene cause autosomal dominant non-syndromic hearing impairment. *Nature genetics* **19**, 60–62 (1998).

42. Michalski, N. & Petit, C. Genetics of auditory mechano-electrical transduction. *Pflügers Archiv-European Journal of Physiology* **467**, 49–72 (2015).

43. Ishikawa, K. et al. A Japanese family showing high-frequency hearing loss with KCNQ4 and TECTA mutations. *Acta oto-laryngologica* **134**, 557–563 (2014).

44. Collin, R. W. et al. Mid-frequency DFNA8/12 hearing loss caused by a synonymous TECTA mutation that affects an exonic splice enhancer. *European Journal of Human Genetics* **16**, 1430–1436 (2008).

45. Moteki, H. et al. TECTA mutations in Japanese with mid-frequency hearing loss affected by zona pellucida domain protein secretion. *Journal of human genetics* **57**, 587–592 (2012).

46. Dunn, O. J. Multiple comparisons among means. *Journal of the American Statistical Association* **56**, 52–64 (1961).

47. Davis, R. W. A review of the multi-level adaptations for maximizing aerobic dive duration in marine mammals: from biochemistry to behavior. *Journal of Comparative Physiology B* **184**, 23–53 (2014).

48. Parker, J. et al. Genome-wide signatures of convergent evolution in echolocating mammals. *Nature* **502**, 228–231 (2013).
49. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).

50. Andrews, S. FQC: A quality control tool for high throughput sequence data. *Reference Source* (2010).

51. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

52. O’Connell, J. et al. NCtrim: optimized trimming of Illumina mate pair reads. *Bioinformatics* **31**, 2035–2037 (2015).

53. Gnerre, S. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences* **108**, 1513–1518 (2011).

54. Peng, Y., Leung, H. C., Yu, S.-M. & Chin, F. Y. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–1428 (2012).

55. Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2010).

56. Mandric, I. & Želikovski, A. ScaffMatch: scaffolding algorithm based on maximum weight matching. *Bioinformatics* **31**, 2632–2638 (2015).

57. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).

58. Smit, A. & Hubley, R. RepeatModeler Open-1.0. Repeat Masker Website (2010).

59. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* **12**, 1269–1276 (2002).

60. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of families and large genomes. *Bioinformatics* **21**, i351–i358 (2005).

61. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **25**, 3839–3840 (1997).

62. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *Nucleic acids research* **34**, W435–W439 (2006).

63. Altschul, S. F. et al. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).

64. Boeckmann, B. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2876–2879 (2007).

65. Talavera, G. & Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17**, 540–552 (2000).

66. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).

67. Simão, F. A., Waterhouse, R. M., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

68. Jones, D. T., Taylor, W. R. & Thornton, J. M. The rapid generation of mutation data matrices from protein sequences. *Bioinformatics* **14**, 405–406 (1998).

69. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879 (2004).

70. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7 (2008).

71. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17**, 540–552 (2000).

72. Keane, T. M., Creevey, C. J., Pentony, M. M., Naughton, T. J. & McInerney, J. O. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *Molecular biology and evolution* **20**, 345–358 (2003).

73. Keane, T., Naughton, T. & McInerney, J. ModelGenerator: amino acid and nucleotide substitution model selection. *Nucleic acids research* **31**, 1100–1103 (2003).

74. Kadavekonda, R. et al. Ensembl 2012. *Genome biology* **13*, R7 (2012).

75. Larkin, M. A. et al. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2876–2879 (2007).

76. Lanave, C. et al. Amino acid sequence evolution. *Journal of molecular biology* **162**, 363–366 (1982).

77. Altschul, S. F. et al. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).

78. Gascuel, O. The tree of life. *Nature* **405**, 866–871 (2000).

79. Lunter, G., et al. FastTree: computing large phylogenies made simple. *Bioinformatics* **25**, 791–796 (2009).

80. Morgenstern, B. et al. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* **13**, 2178–2185 (2003).

81. Lunter, G., et al. FastTree: computing large phylogenies made simple. *Bioinformatics* **25**, 791–796 (2009).

82. Ayres, C. A. et al. Genome assembly and annotation of the drosophila genome. *Science* **317**, 1572 (2007).

83. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* **34**, W609–W612 (2006).

84. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution* **24**, 1586–1591 (2007).

85. Zhang, J. et al. FASTA: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic acids research* **40**, W580–W584 (2012).

86. Acevedo, B. et al. Aligning multiple genomes with the threaded blockset aligner. *Genome biology* **14**, 708–715 (2004).

87. Marchler-Bauer, A. et al. CD-HIT: NCBI’s conserved domain database. *Nucleic acids research* **43**, D222–D226 (2014).

88. Pöldner, E. et al. Ensembl 2012. *Nucleic acids research* **40**, D84–D90 (2011).

89. Dennis, G. et al. DAVID: database for annotation, visualization, and integrated discovery. *Genome biology* **4**, R60 (2003).
