nature genetics

OPEN

Minke whale genome and aquatic adaptation in cetaceans

Hyung-Soon Yim[1,24], Yun Sung Cho[2,24], Xuanmin Guang[3,24], Sung Gyun Kang[1,4], Jae-Yeon Jeong[1,4], Sun-Shin Cha[1,4,5], Hyun-Myung Oh[1], Jae-Hak Lee[1], Eun Chan Yang[1], Kae Kyoung Kwon[1,4], Yun Jae Kim[1], Tae Wan Kim[1], Wonduck Kim[1], Jeong Ho Jeon[1], Sang-Jin Kim[1,4], Dong Han Choi[1], Sungwoong Jho[2], Hak-Min Kim[2], Junsu Ko[6], Hyunmin Kim[6], Young-Ah Shin[2], Hyun-Ju Jung[6], Yuan Zheng[3], Zhuo Wang[3], Yan Chen[3], Ming Chen[3], Awei Jiang[3], Erli Li[3], Shu Zhang[3], Haolong Hou[7], Tae Hyung Kim[6], Lili Yu[3], Sha Liu[3], Kung Ahn[6], Jesse Cooper[6], Sin-Gi Park[6], Chang Pyo Hong[6], Wook Jin[8], Heui-Soo Kim[9], Chankyu Park[10], Kyooyeol Lee[10], Sung Chun[11], Phillip A Morin[12], Stephen J O'Brien[13], Hang Lee[14], Jumpei Kimura[15], Dae Yeon Moon[16], Andrea Manica[17], Jeremy Edwards[18], Byung Chul Kim[2], Sangsoo Kim[19], Jun Wang[3,20,21], Jong Bhak[2,6,22,23], Hyun Sook Lee[1,4] & Jung-Hyun Lee[1,4]

The shift from terrestrial to aquatic life by whales was a substantial evolutionary event. Here we report the whole-genome sequencing and *de novo* assembly of the minke whale genome, as well as the whole-genome sequences of three minke whales, a fin whale, a bottlenose dolphin and a finless porpoise. Our comparative genomic analysis identified an expansion in the whale lineage of gene families associated with stress-responsive proteins and anaerobic metabolism, whereas gene families related to body hair and sensory receptors were contracted. Our analysis also identified whale-specific mutations in genes encoding antioxidants and enzymes controlling blood pressure and salt concentration. Overall the whale-genome sequences exhibited distinct features that are associated with the physiological and morphological changes needed for life in an aquatic environment, marked by resistance to physiological stresses caused by a lack of oxygen, increased amounts of reactive oxygen species and high salt levels.

Cetaceans include whales, dolphins and porpoises. They are placed phylogenetically in Cetartiodactyla, the clade that includes Cetacea and Artiodactyla (even-toed ungulates such as the hippopotamus, cow and pig)[1]. Whales and modern terrestrial artiodactyls are related to *Indohyus* (an extinct semiaquatic deer-like ungulate), from which they are known to have split 54 million years ago[2]. Underwater adaptations of cetaceans to physiological stress, along with their unique morphology, are interesting. The minke whale is the most abundant baleen whale and is classified into two species: the common minke whale (*Balaenoptera acutorostrata*) and the Antarctic minke whale (*Balaenoptera bonaerensis*)[3]. The wide geographical distribution of the minke whale makes it an ideal candidate for whole-genome sequencing. In addition to a low-coverage (2.59×) assembly of the bottlenose dolphin (*Tursiops truncatus*) genome[4–6], there are now several sequenced cetaceans that can be used as resources for evolutionary and population-management studies. We report the *de novo* assembly of the common minke whale genome and a comparative analysis of additional genomic sequences (~30× depth, aligned to the reference genomes but not assembled) of three minke whales, a fin whale (*Balaenoptera physalus*), a bottlenose dolphin and a finless porpoise (*Neophocaena phocaenoides*) (**Supplementary Tables 1** and **2**).

We extracted DNA from male minke whale muscle and sequenced it to a 128× average depth of coverage using the Illumina HiSeq 2000

[1]Korea Institute of Ocean Science and Technology, Ansan, Republic of Korea. [2]Personal Genomics Institute, Genome Research Foundation, Suwon, Republic of Korea. [3]Beijing Genomics Institute (BGI)-Shenzhen, Shenzhen, China. [4]Department of Marine Biotechnology, University of Science and Technology, Daejeon, Republic of Korea. [5]Ocean Science and Technology School, Korea Maritime University, Busan, Republic of Korea. [6]Theragen BiO Institute, TheragenEtex, Suwon, Republic of Korea. [7]Shaanxi Yulin Energy Group Co. Ltd., Yulin, Shaanxi, China. [8]Department of Molecular Medicine, School of Medicine, Gachon University, Incheon, Republic of Korea. [9]Department of Biological Sciences, College of Natural Sciences, Pusan National University, Busan, Republic of Korea. [10]Laboratory of Genome Biology, Department of Animal Biotechnology, Konkuk University, Seoul, Republic of Korea. [11]Division of Genetics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [12]Marine Mammal and Turtle Division, Southwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, La Jolla, California, USA. [13]Theodosius Dobzhansky Center for Genome Bioinformatics, St. Petersburg State University, St. Petersburg, Russia. [14]College of Veterinary Medicine, Seoul National University, Seoul, Republic of Korea. [15]Department of Anatomy and Cell Biology, College of Veterinary Medicine, Seoul National University, Seoul, Republic of Korea. [16]Marine Biodiversity Institute of Korea (MABIK), Ministry of Ocean and Fisheries, Sejong, Republic of Korea. [17]Evolutionary Ecology Group, Department of Zoology, University of Cambridge, Cambridge, UK. [18]Department of Molecular Genetics and Microbiology, University of New Mexico Health Sciences Center, Albuquerque, New Mexico, USA. [19]School of Systems Biomedical Science, Soongsil University, Seoul, Republic of Korea. [20]Department of Biology, University of Copenhagen, Copenhagen, Denmark. [21]King Abdulaziz University, Jeddah, Saudi Arabia. [22]Program in Nano Science and Technology, Department of Transdisciplinary Studies, Seoul National University, Suwon, Republic of Korea. [23]Advanced Institutes of Convergence Technology Nano Science and Technology, Suwon, Republic of Korea. [24]These authors contributed equally to this work. Correspondence should be addressed to Jung-Hyun Lee (jlee@kiost.ac), H.S.L. (leeh522@kiost.ac) or J.B. (jongbhak@genomics.org).

Received 18 June; accepted 1 November; published online 24 November 2013; doi:10.1038/ng.2835

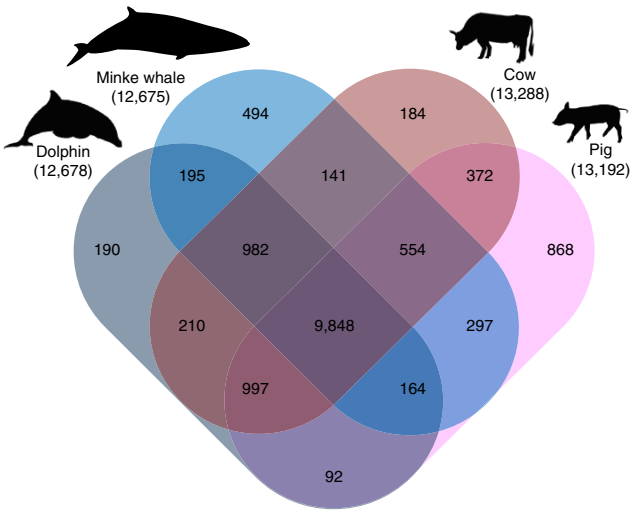

Figure 1 Orthologous gene clusters in the artiodactyl lineage. Shown is a Venn diagram of unique and shared gene families in the minke whale, bottlenose dolphin, cow and pig genomes. The total numbers of gene families are given in parentheses.

platform (**Supplementary Tables 3–5**). Raw reads were assembled into 104,325 scaffolds totaling 2.44 Gb in length (**Supplementary Figs. 1–4** and **Supplementary Tables 6–9**). We assessed the quality of the assembly by aligning the assembled minke whale transcripts onto the scaffolds (>98% coverage) and by using a core eukaryotic gene mapping method[7] (>98.6% conserved genes) (**Supplementary Tables 10–13**). Additionally, we validated heterozygous single-nucleotide variants (SNVs) by Sanger sequencing (**Supplementary Fig. 5** and **Supplementary Tables 14** and **15**). We identified all four analyzed minke whales as North Pacific minke whales (*B. acutorostrata scammoni*) by mapping their raw reads to a previously published mitochondrial genome[8] (**Supplementary Figs. 6** and **7**). Minke whales have 21 pairs of autosomes and a pair of sex chromosomes ($2n = 44$), which is common in cetacean[9]. We could identify eight scaffolds as a small fraction of sex chromosomes (**Supplementary Table 16**).

We found that the minke whale genome contains 20,605 genes (**Supplementary Tables 17–19**) and 2,598 noncoding RNAs (**Supplementary Table 20**). Repetitive elements occupy 37.3% of the whole genome (**Supplementary Figs. 8** and **9** and **Supplementary Tables 21–24**). We confirmed genes on the basis of the transcriptomes of eight organs sequenced from an additionally acquired minke whale sample (**Supplementary Table 10**). We constructed orthologous gene clusters using eight mammalian genomes (**Supplementary Table 25**). We found that the minke whale genome contains 12,675 orthologous gene families, excluding singletons, 9,848 of which are shared by all four artiodactyl genomes (minke whale, bottlenose dolphin, cow and pig). Of these gene families, 494 are specific to the minke whale (**Fig. 1**; estimates of divergence time are shown in **Supplementary Figs. 10** and **11**). Additionally, we estimated segmental duplication (33.4 Mb) and genomic synteny (30–45%) in the minke whale genome (**Supplementary Figs. 12** and **13**, **Supplementary Tables 26–29** and **Supplementary Note**). An inspection of the gene families showed that olfactory, rhodopsin-like G protein–coupled receptor and mammalian taste receptor domains were markedly under-represented in the whales compared to in the cow and pig (**Supplementary Tables 30** and **31**). We further analyzed the genome to identify all olfactory receptor genes and found that the number of these genes is much lower in whales than in other mammals (**Supplementary Figs. 14** and **15**, **Supplementary Tables 32** and **33** and **Supplementary Note**).

We investigated the genotypes underlying the marine adaptations of the whale lineage by analyzing the expansion or contraction of gene families, species-specific amino acid changes and positively selected genes (PSGs). We found that the minke whale genome contains 1,156 expanded and 2,048 contracted gene families (**Fig. 2a** and **Supplementary Tables 34** and **35**). Compared with other non-whale mammals, the whale lineage contains a total of 4,773 genes with unique amino acid changes (fixed in the four minke whales and two bottlenose dolphins), and 574 genes had minke whale–specific amino acid changes (fixed only in the four minke whales). Of the 4,773 genes, 695 encoded function-altering amino acid changes that were specific to the whale lineage (**Supplementary Table 36**). We identified PSGs, on the basis of d_N/d_S ratios (nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site), by comparing the whale genomes with those of cow and pig using the branch-site likelihood ratio test[10]. We identified 279 and 557 PSGs in the minke whale and bottlenose dolphin, respectively, whereas 64 PSGs were present in both (**Supplementary Tables 37–43**). Additionally, we identified rapidly evolving gene ontology (GO) categories[11] in the minke whale and bottlenose dolphin (**Supplementary Tables 44** and **45**), as well as copy number variations in the fin whale and finless porpoise (**Supplementary Tables 46–48** and **Supplementary Note**).

Notably, a number of whale-specific genes were strongly associated with stress resistance. The peroxiredoxin (*PRDX*) family, which has an important role in eliminating peroxides and in redox signaling generated during metabolism[12,13], was markedly expanded (GO:0051920, $P = 0.000030$, Fisher's exact test, seven genes; **Supplementary Table 34**), meaning there was an increase in gene number in the whale lineage (**Fig. 2b** and **Supplementary Tables 49** and **50**). *PRDX1* was expanded in the minke whale (five copies) and bottlenose dolphin (two copies). The fin whale and finless porpoise also had expanded *PRDX1* homolog genes. Furthermore, *PRDX3* was expanded in the two baleen whales (two copies), and *PRDX4* was positively selected in the minke whale and bottlenose dolphin.

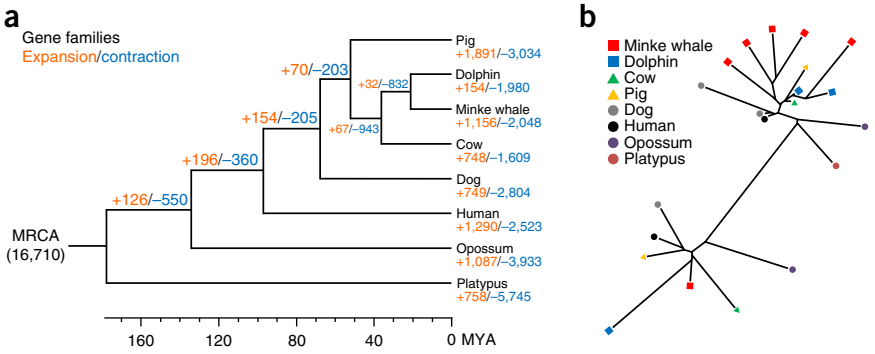

Figure 2 Relationship of the minke whale to other mammalian species. (**a**) Gene family expansion or contraction. The numbers indicate the number of gene families that have expanded (orange) or contracted (blue) since the split from a common ancestor. MYA, million years ago; MRCA, most recent common ancestor. Timelines indicate the divergence times among species. (**b**) The expanded peroxiredoxin (*PRDX*) gene family in the whale lineage.

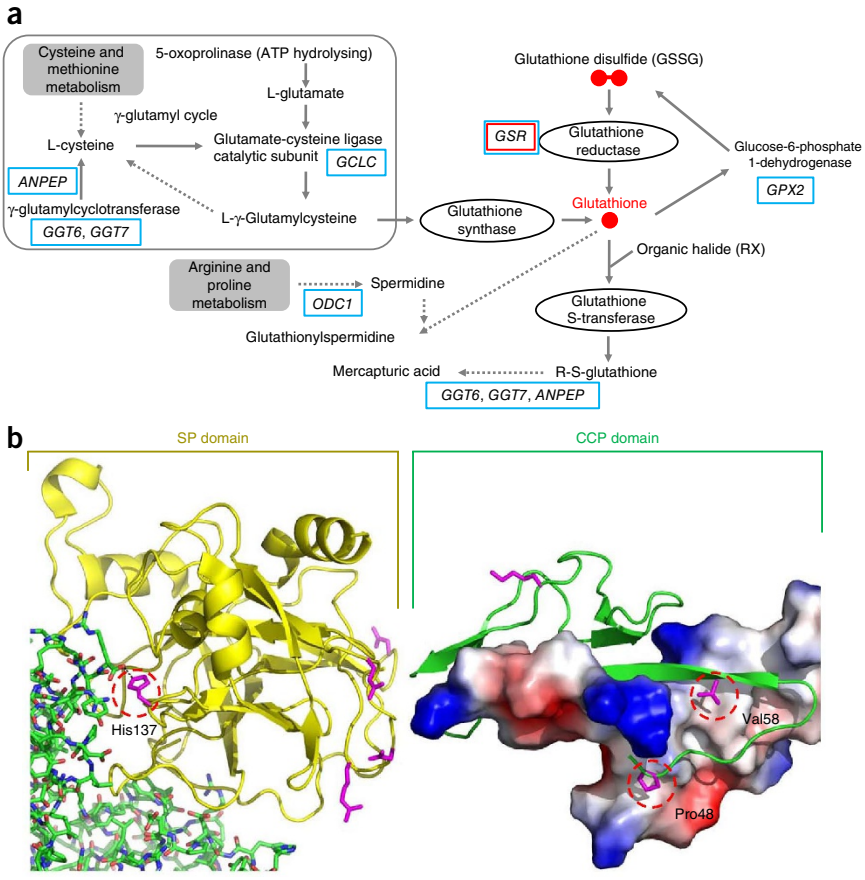

Figure 3 Cetacean-specific amino acid changes in glutathione metabolism–associated genes and haptoglobin. (**a**) A positively selected gene (*GSR*) in the bottlenose dolphin is shown in a red rectangle. Genes with cetacean-specific amino acid changes (*GSR*, *GPX2*, *GGT6*, *GGT7*, *ANPEP*, *ODC1* and *GCLC*) are shown in blue rectangles. The seven cetacean-specific genes are involved in glutathione metabolism pathways (KEGG pathway map00480). The solid lines indicate direct relationships between enzymes and metabolites. The dashed lines indicate that more than one step is involved in a process. (**b**) The positions of unique amino acid changes in the crystal structure of the haptoglobin-hemoglobin complex. The haptoglobin protein is shown in a cartoon form; the CCP domain is green, and the SP domain is yellow. Of the ten amino acid changes, eight positions are represented by violet sticks; the other two positions are not displayed because they were not included in the complex structure. Hemoglobin is shown with green sticks, and the CCP domain of the contacting haptoglobin is represented in an electrostatic potential surface model (blue, positive; red, negative; white, neutral). The black dots indicate the polar interaction between His137 of haptoglobin and the terminal carboxylate of hemoglobin.

The level of O-linked N-acetylglucosaminylation (O-GlcNAcylation) in numerous nucleocytoplasmic proteins is known to increase in response to multiple forms of cellular stress, such as hypoxia, oxidative stress and osmotic stress[14–16]. O-GlcNAc transferase (encoded by *OGT*), the enzyme that can catalyze the addition of a single N-acetylglucosamine to a serine or threonine residue through an O-glycosidic linkage, was expanded in the minke whale (3 copies) and in the bottlenose dolphin (11 copies), whereas the cow and pig had only 1 copy each (**Supplementary Fig. 16** and **Supplementary Tables 49** and **50**). The fin whale and finless porpoise also had expanded *OGT* homolog genes.

Perhaps the most marked environmental adaptation for a whale is deep diving, which induces hypoxia. Under hypoxic conditions, reactive oxygen species (ROS) are generated by several cellular mechanisms[17,18]. Glutathione is a well-known antioxidant that prevents damage to important cellular components by ROS[19]. Seven glutathione metabolism pathway genes (*GPX2*, *ODC1*, *GSR*, *GGT6*, *GGT7*, *GCLC* and *ANPEP*) showed cetacean-specific amino acid changes; these changes were present in the four minke whales, a fin whale, two bottlenose dolphins and a porpoise (**Fig. 3a** and **Supplementary Figs. 17–23**). *GSR* in the glutathione metabolism pathway was also positively selected in the dolphin. It is known that the increased expression of *GSR* increases the antioxidant capacity of cells[20]. Furthermore, functional categories, such as antioxidant activity (GO:0016209, $P = 0.010$, 13 genes) and oxidoreductase activity (GO:0016491, $P = 0.00000035$, 162 genes), were enriched in the minke whale genome (**Supplementary Table 34**). These signatures likely reflect adaptation to increased diving duration, as these genes can combat the damaging effects of hypoxia-induced ROS. To test this hypothesis, we measured glutathione levels experimentally. Cultured kidney cells from the Atlantic spotted dolphin (*Stenella frontalis*) showed an increased ratio of reduced glutathione to glutathione disulfide when subjected to hypoxic or oxidative stress (**Supplementary Fig. 24** and **Supplementary Note**).

Haptoglobin, an antioxidant protein that functions by controlling heme-induced ROS, exhibited ten cetacean-specific amino acid changes (**Supplementary Figs. 25** and **26**). Haptoglobins bind free plasma hemoglobins, thereby preventing the loss of iron through the kidneys and protecting against renal damage caused by hemoglobin-derived ROS[21]. We identified two genetic variations (encoding p.Pro48Leu and p.Val58Leu) at the dimeric interface between the complement control protein (CCP) domains in two residues that are in hydrophobic contact (**Fig. 3b**). In these cases, replacement with bulkier hydrophobic residues appears to strengthen the contact. We observed another notable variant (p.His137Asn) on the hemoglobin-interacting face of the serine protease (SP) domain. His137 participates in a polar interaction with the C-terminal carboxylate of hemoglobin. The p.His137Asn substitution could facilitate two polar interactions between the amide side chain of asparagine and the terminal carboxylate, thereby strengthening the interaction between hemoglobin and haptoglobin.

In whales, blood lactate concentration increases after prolonged diving[22,23], and hypoxia is known to control lactate concentration by activating hypoxia-inducible factor[24]. Lactate dehydrogenase (encoded by *LDH*) is the enzyme responsible for converting pyruvate to lactate. In our analysis, we found that the *LDHA* homolog genes had undergone an expansion in mammals that are known to live under hypoxic conditions, namely whales and naked mole rats (**Supplementary Fig. 27**). Additionally, we discovered that the genes encoding homologs of monocarboxylate transporter 1 (encoded by *SLC16A1*, also called *MCT1*), which catalyze the rapid transport of

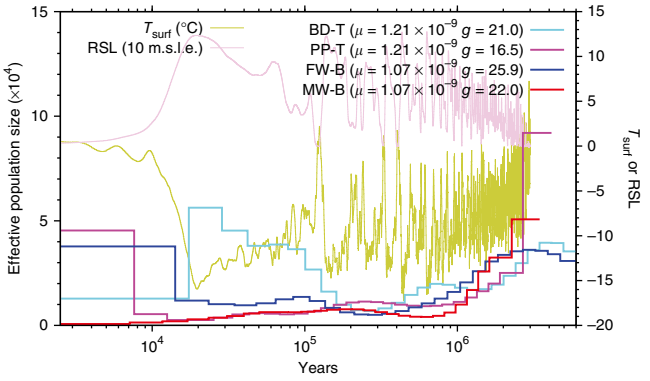

Figure 4 Estimated whale population size history. T_{surf}, atmospheric surface air temperature; RSL, relative sea level; 10 m.s.l.e., 10 m sea level equivalent; MW, minke whale; FW, fin whale; BD, bottlenose dolphin; PP, finless porpoise; *g*, generation time; μ, mutation rate (per site, per year). Minke whale and fin whale data were generated on the basis of comparisons with minke whale scaffolds ("-B" after the species abbreviation) during SNV calling, whereas the bottlenose dolphin and finless porpoise data were generated on the basis of comparisons with the bottlenose dolphin scaffolds ("-T" after the species abbreviation) during SNV calling.

monocarboxylates such as lactate and pyruvate across the plasma membrane, had also expanded in the whale lineage and in naked mole rats (**Supplementary Fig. 28**).

Whales extract the majority of their water from food by metabolizing fat, but they still consume seawater under certain circumstances[25]. The renin-angiotensin-aldosterone system (RAAS) is a key hormone system that regulates blood pressure and water balance in response to sodium level. Notably, we found functional changes in angiotensin-converting enzyme 2 (encoded by *ACE2*) in the whale lineage (p.Val747Ala, p.Asp798Gly and p.Gln801His in minke and fin whales and p.Asp784Gly in bottlenose dolphin and finless porpoise; **Supplementary Fig. 29**), and five genes in the RAAS pathway (*AGTR1*, *ANPEP*, *LNPEP*, *MME* and *THOP1*; **Supplementary Figs. 23** and **30–33**) had cetacean-specific amino acid changes.

Mysticeti whale species, including minke whales, grow baleen instead of teeth. We observed that *ENAM*, *MMP20* and *AMEL*, which are involved in tooth enamel formation and biomineralization[26,27], are pseudogenes with premature stop codons in the baleen whales (**Supplementary Figs. 34–37**). Keratin-related gene families, which are essential for hair formation, were contracted specifically in the whale lineage (**Supplementary Fig. 38**). Additionally, several *HOX* genes (*HOXA5*, *HOXB1*, *HOXB2*, *HOXB5*, *HOXD12* and *HOXD13*), which have an important role in the body plan and embryonic development[28], were positively selected in the whale lineage compared to terrestrial mammals, reflecting the morphological adaptation of the whale to the aquatic environment (**Supplementary Fig. 39** and **Supplementary Note**). The adaptations of whales to echolocation (**Supplementary Fig. 40**), blood clotting (**Supplementary Table 51**) and oxygen transportation (**Supplementary Fig. 41** and **Supplementary Tables 52–54**) are described in the **Supplementary Note**.

Analysis of whole-genome sequences can provide a general overview of the total genetic variation in a species, and in this study we identified 1.37–1.59 million heterozygous SNVs in the genomes of the three resequenced minke whales (**Supplementary Tables 55** and **56**). This gave an estimated nucleotide diversity (mean per-nucleotide heterozygosity) of 0.00061, which is comparable to the nucleotide diversity observed in humans (0.00069)[29]. The nucleotide diversities of the fin whale (0.00151) and bottlenose dolphin (0.00142) were higher than those of the minke whale and finless porpoise (0.00086). Although blue and fin whales are as genetically distant as gorillas and humans[30], they are known to interbreed and produce hybrid individuals[31], which could explain the observed high level of heterozygosity in the fin whale. Similarly, bottlenose dolphins are known to hybridize with other dolphins[32–34]. We also inferred a marked population bottleneck in the demographic history of the whale using the pairwise sequentially Markovian coalescent (PSMC) model[35] (**Fig. 4** and **Supplementary Table 57**). The minke whale and finless porpoise genomes indicated no substantial population increase during the upper Pleistocene age (12,000–130,000 years ago), whereas the fin whale and bottlenose dolphin populations increased, suggesting that either a population expansion or substantial genetic exchanges through hybridization occurred.

To the best of our knowledge, the minke whale reference genome is the first high-depth marine mammalian genome to be sequenced. The cetacean genomes support hypotheses regarding adaptation to hypoxic resistance, metabolism under limited oxygen and high-salt conditions and the development of unique morphological traits. In particular, the expansion of antioxidant-related genes and whale-specific variations in glutathione-associated and haptoglobin proteins are evidence for adaptation to hypoxic conditions during diving. These data will contribute to future studies of marine mammal diseases, conservation and evolution.

URLs. Minke whale genome, http://whalegenome.net/; SOAP, http://soap.genomics.org.cn/; Ensembl, http://www.ensembl.org/index.html; KEGG, http://www.genome.jp/kegg/; Repbase, http://www.girinst.org/repbase/index.html; RepeatMasker, http://repeatmasker.org/; MCMCtree, http://abacus.gene.ucl.ac.uk/software/paml.html.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. The minke whale whole-genome shotgun project has been deposited at the DNA Data Bank of Japan, European Molecular Biology Laboratory and GenBank under accession ATDI00000000. The version described in this paper is the first version, ATDI01000000. Raw DNA and RNA sequencing reads have been submitted to the NCBI Sequence Read Archive database (SRA090057 and SRA091100).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

No animals were killed or captured as a result of these studies. Samples of four minke whales and a finless porpoise for sequencing were acquired from the east coast of Korea after they were accidently killed and investigated by the maritime police. The sample from a bottlenose dolphin was obtained from Marine Park in Jeju Island, Korea. The fin whale sample was collected from a dead stranded fin whale by the Southwest Fisheries Science Center (SWFSC 134239) under US Marine Mammal Permit 14097-01, and the sample was transported internationally under a Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES) permit. We thank C.-W. Kim (Marine Park Co., Ltd.) and Y.-R. An (Cetacean Research Institute) for the bottlenose dolphin and minke whale samples, respectively. We also thank M. Bhak for editing. We thank many people not listed as authors who provided feedback, samples and encouragement, especially the Cetacean Research Institute in Ulsan, Korea. This work was supported by the Korea Institute of Ocean Science and Technology (KIOST) in-house program (PE98993), the Marine and Extreme Genome Research Center program of the Ministry of Oceans and Fisheries, Korea. This work was also supported by the Industrial Strategic Technology Development Program 10040231,

Bioinformatics Platform Development for Next-Generation Bioinformation Analysis. The Russian Ministry of Science supported S.J.O. (mega-grant 11.G34.31.0068).

AUTHOR CONTRIBUTIONS

The whale genome project was initiated by KIOST. Research collaboration and analysis was carried out by KIOST, the Genome Research Foundation, BGI and the institutes of other participating authors within the whale genome consortium. Jung-Hyun Lee, H.S.L. and J.B. supervised the project. H.-S.Y. coordinated the project. P.A.M., H.L., J. Kimura and D.Y.M. provided samples, advice and associated information. Library construction, sequencing and genome assembly for the draft reference genome were carried out by Y.Z., E.L. and S.Z. Several cetacean genome resequencings were performed by H.-J.J. Experimental validations were performed by S.G.K., W.J. and K.A. Bioinformatics data processing and analyses of genetic variation data were carried out by X.G., Z.W., Y.C., H.H., M.C., A.J., L.Y., S.L., Y.S.C., J.-Y.J., S.-S.C., H.-M.O., Jae-Hak Lee, E.C.Y., K.K.K., Y.J.K., T.W.K., W.K., J.H.J., S.-J.K., D.H.C., S.J., H.-M.K., J. Ko, H.K., T.H.K., J.C., S.-G.P., Y.-A.S., C.P.H., S.C., H.-S.K., K.L. and C.P. H.-S.Y., Y.S.C., X.G., Z.W., S.G.K., J.-Y.J., S.-S.C., H.-M.O., Jae-Hak Lee, E.C.Y., K.K.K., Y.J.K., T.W.K., W.K., J.H.J., S.-J.K., J.W., S.J.O., A.M., J.E., S.K., B.C.K., H.S.L., Jung-Hyun Lee and J.B. wrote, edited and revised the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Thewissen, J.G., Cooper, L.N., Clementz, M.T., Bajpai, S. & Tiwari, B.N. Whales originated from aquatic artiodactyls in the Eocene epoch of India. *Nature* **450**, 1190–1194 (2007).
2. Dawkins, R. *The Ancestor's Tale, A Pilgrimage to the Dawn of Life* (Houghton Mifflin Harcourt Press, Boston, 2004).
3. Wilson, D.E. & Reeder, D.M. *Mammal Species of the World* (Johns Hopkins University Press, 2005).
4. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
5. McGowen, M.R., Grossman, L.I. & Wildman, D.E. Dolphin genome provides evidence for adaptive evolution of nervous system genes and a molecular rate slowdown. *Proc. Biol. Sci.* **279**, 3643–3651 (2012).
6. Sun, Y.B. *et al.* Genome-wide scans for candidate genes involved to the aquatic adaptation of dolphins. *Genome Biol. Evol.* **5**, 130–139 (2013).
7. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).
8. Arnason, U., Gullberg, A. & Janke, A. Mitogenomic analyses provide new insights into cetacean origin and evolution. *Gene* **333**, 27–34 (2004).
9. O'Brien, S.J., Menninger, J.C. & Nash, W.G. *An Atlas of Mammalian Genomes* (John Wiley & Sons Publishers, New York, 2006).
10. Nielsen, R. *et al.* A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol.* **3**, el70 (2005).
11. Qiu, Q. *et al.* The yak genome and adaptation to life at high altitude. *Nat. Genet.* **44**, 946–949 (2012).
12. Rhee, S.G., Chae, H. & Kim, K. Peroxiredoxins: a historical overview and speculative preview of novel mechanisms and emerging concepts in cell signaling. *Free Radic. Biol. Med.* **38**, 1543–1552 (2005).
13. Neumann, C.A., Cao, J. & Manevich, Y. Peroxiredoxin 1 and its role in cell signaling. *Cell Cycle* **8**, 4072–4078 (2009).
14. Zachara, N.E. *et al.* Dynamic O-GlcNAc modification of nucleocytoplasmic proteins in response to stress. A survival response of mammalian cells. *J. Biol. Chem.* **279**, 30133–30142 (2004).
15. Ngoh, G.A., Watson, L.J., Facundo, H.T. & Jones, S.P. Augmented O-GlcNAc signaling attenuates oxidative stress and calcium overload in cardiomyocytes. *Amino Acids* **40**, 895–911 (2011).
16. Jones, S.P. *et al.* Cardioprotection by N-acetylglucosamine linkage to cellular proteins. *Circulation* **117**, 1172–1182 (2008).
17. Gonchar, O. & Mankovska, I. Antioxidant system in adaptation to intermittent hypoxia. *J. Biol. Sci.* **10**, 545–554 (2010).
18. Blokhina, O., Virolainen, E. & Fagerstedt, K.V. Anioxidants, oxidative damage and oxygen deprivation stress: a review. *Ann. Bot.* (*Lond.*) **91**, 179–194 (2003).
19. Pompella, A., Visvikis, A., Paolicchi, A., De Tata, V. & Casini, A.F. The changing faces of glutathione, a cellular protagonist. *Biochem. Pharmacol.* **66**, 1499–1503 (2003).
20. Foyer, C.H. *et al.* Overexpression of glutathione reductase but not glutathione synthetase leads to increases in antioxidant capacity and resistance to photoinhibition in Poplar trees. *Plant Physiol.* **109**, 1047–1057 (1995).
21. Andersen, C.B. *et al.* Structure of the haptoglobin-haemoglobin complex. *Nature* **489**, 456–459 (2012).
22. Shaffer, S.A., Costa, D.P., Williams, T.M. & Ridgway, S.H. Diving and swimming performance of white whales, *Delphinapterus leucas*: an assessment of plasma lactate and blood gas levels and respiratory rates. *J. Exp. Biol.* **200**, 3091–3099 (1997).
23. Williams, T.M., Haun, J.E. & Friedl, W.A. The diving physiology of bottlenose dolphins (*Tursiops truncatus*). I. Balancing the demands of exercise for energy conservation at depth. *J. Exp. Biol.* **202**, 2739–2748 (1999).
24. Firth, J.D., Ebert, B.L. & Ratcliffe, P.J. Hypoxic regulation of lactate dehydrogenase A. Interaction between hypoxia-inducible factor 1 and cAMP response elements. *J. Biol. Chem.* **270**, 21021–21027 (1995).
25. Ortiz, R.M. Osmoregulation in marine mammals. *J. Exp. Biol.* **204**, 1831–1844 (2001).
26. Rajpar, M.H., Harley, K., Laing, C., Davies, R.M. & Dixon, M.J. Mutation of the gene encoding the enamel-specific protein, enamelin, causes autosomal-dominant amelogenesis imperfecta. *Hum. Mol. Genet.* **10**, 1673–1677 (2001).
27. Meredith, R.W., Gatesy, J., Cheng, J. & Springer, M.S. Pseudogenization of the tooth gene enamelysin (*MMP20*) in the common ancestor of extant baleen whales. *Proc. Biol. Sci.* **278**, 993–1002 (2011).
28. Liang, L. *et al.* Adaptive evolution of the *Hox* gene family for development in bats and dolphins. *PLoS ONE* **8**, e65944 (2013).
29. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
30. Arnason, U. & Gullberg, A. Comparison between the complete mtDNA sequences of the blue and fin whale, two species that can hybridize in nature. *J. Mol. Evol.* **37**, 312–322 (1993).
31. Palumbi, S.R. & Cipriano, F. Species identification using genetic tools: the value of nuclear and mitochondrial gene sequences in whale conservation. *J. Hered.* **89**, 459–464 (1998).
32. Reeves, R., Stewart, B., Clapham, P. & Powell, J. *National Audubon Society: Guide to Marine Mammals of the World* (Alfred A. Knopf, New York, 2002).
33. Caballero, S. & Baker, C.S. Captive-born intergeneric hybrid of a Guiana and bottlenose dolphin: *Sotalia guianensis* × *Tursiops truncatus*. *Zoo Biol.* **29**, 647–657 (2010).
34. Herzing, D., Moewe, K. & Brunnick, B. Interspecies interactions between Atlantic spotted dolphins, *Stenella frontalis* and bottlenose dolphins, *Tursiops truncatus*, on Great Bahama Bank, Bahamas. *Aquat. Mamm.* **29**, 335–341 (2003).
35. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).

ONLINE METHODS

Genome sequencing and assembly. The samples used for genome sequencing were acquired from four minke whales and a finless porpoise that had been accidently killed near the east coast of Korea; these incidents were investigated by the Korean maritime police. The bottlenose dolphin sample was obtained from Marine Park in Jeju Island, Korea. The fin whale sample was collected from a dead stranded fin whale by the Southwest Fisheries Science Center. Libraries with different insert sizes were constructed at BGI-Shenzhen (import permit: APO/IL 378/12; export permit: ES2012-00776). The insert sizes of the libraries were 170 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb and 20 kb. The libraries were sequenced using a HiSeq2000 instrument. Other cetacean species (three additional minke whales, one fin whale, one bottlenose dolphin and one finless porpoise) were sequenced at the Theragen BiO Institute (TBI), Korea, using a HiSeq2000 instrument with a 400-bp insert library. We applied filtering criteria to reduce the effects of sequencing errors in the assembly (**Supplementary Note**).

The corrected reads were used to complete the genome assembly using SOAPdenovo-1.05 (ref. 36). Only qualified data were used in the genome assembly. First, the short insert size library data were used to construct a de Bruijn graph. The tips, merged bubbles and connections with low coverage were removed before resolving the small repeats. Second, all qualified reads were realigned with the contig sequences. The number of shared paired-end relationships between pairs of contigs was calculated and weighted with the rate of consistent and conflicting paired ends before constructing the scaffolds in a stepwise manner from the short–insert size paired ends to the long–insert size paired ends. Third, the gaps between the constructed scaffolds were composed mainly of repeats, which were masked during scaffold construction. These gaps were closed using the paired-end information to retrieve read pairs in which one end mapped to a unique contig and the other was located in the gap region. Subsequently, local assembly was conducted for these collected reads. SSPACE v1-1 (ref. 37) was also used to build the scaffolds. The assembly quality was assessed by mapping the DNA and short RNA reads to the scaffolds; 91.0% of the DNA reads and 78.9% of the RNA reads were mapped (**Supplementary Tables 9** and **10**). The mapping was conducted using BWA-0.6.2 (ref. 38), and SNVs and small insertions or deletions (indels) were called using SAMtools-0.1.18 (ref. 39) with the default options, except that the '–d 5 -D 150' options were used when 'vcfutils.pl varFilter' was executed to filter the SNVs and indels. A total of 141 heterozygous SNVs, which were located in nonrepeat regions, were validated using the Sanger sequencing method, and 139 (98.6%) were true heterozygous SNVs (**Supplementary Fig. 5** and **Supplementary Table 15**). In order to assess the assembly quality, the eight minke whale transcriptomes were assembled using Trinity[40], and all transcripts were longer than 200 bp. The assembled transcripts were aligned to the minke whale scaffolds using BLAT[41] with default options except for an identity cutoff of 90%. We found that over 98% of the assembled transcripts were covered by the minke whale scaffolds (**Supplementary Table 12**). Additionally, we used a core eukaryotic gene mapping analysis CEGMA method[7] to identify the core genes in the minke whale genome assembly. A total of 452 core eukaryotic genes (98.69%) out of 458 were found in the minke whale assembly (**Supplementary Table 13**).

Genome annotation. The genome was searched for repetitive elements using Tandem Repeats Finder[42] version 4.04. Transposable elements were identified using homology-based approaches. The Repbase (version 16.10) database of known repeats and a *de novo* repeat library generated by RepeatModeler[43] were used. This database was used to find repeats with software such as RepeatMasker version 3.3.0. Four types of noncoding RNAs (microRNAs, transfer RNAs, ribosomal RNAs and small nuclear RNAs) were also annotated using tRNAscan-SE[44] (version 1.23) and the Rfam database[45] (Release 9.1) (**Supplementary Note**).

The locations and structures of genes, as well as their biological functions and pathways, were predicted using three approaches (**Supplementary Tables 17** and **18**). First, *de novo* prediction was performed using the repeat-masked genome based on a hidden Markov model. The programs used were AUGUSTUS (version 2.5.5)[46] and GENSCAN (version 1.0)[47]. Second, homologous proteins in other species (from the Ensembl 64 release) were mapped to the genome using tBLASTn (Blast 2.2.23)[48] with an *E*- value cutoff of 1×10^{-5}. The aligned sequence and its query proteins were then filtered and passed to GeneWise (version 2.2.0)[49] to search for accurately spliced alignments. Third, source evidence generated using the above two approaches was integrated with GLEAN-1-0-1 (ref. 50) to produce a consensus gene set. Additionally, transcriptome sequencing data were mapped to the genome using TopHat[51], and gene models were predicted using Cufflinks[52]. Then, additional gene models were added (mainly from the Cufflinks predictions) to the GLEAN gene set to construct the final gene set. Gene functions were assigned on the basis of the best matches in the alignments using BLASTP with the SwissProt and TrEMBL databases (Uniprot release 2011-08)[53,54]. The gene motifs and domains were determined using InterProScan (version 4.7)[55] against public protein databases, including ProDom[56], PRINTS[57,58], Pfam[59], SMART[60], PANTHER[61] and PROSITE[62]. The GO[63] IDs of each gene were obtained from the corresponding InterPro entries. All genes were aligned against KEGG[64] genes (release 58) (**Supplementary Table 19**).

Gene families. Orthologous gene sets were used for genome comparisons. The TreeFam methodology[65] was used to define a gene family, which represents a group of genes descended from a single gene in the last common ancestor. BLASTP was applied to all protein sequences using a database containing a protein data set from all species with *E* values $<1 \times 10^{-7}$, and fragmental alignments were conjoined for each gene pair by Solar. A connection (edge) was assigned between two nodes (genes) if >1/3 of the region aligned to both genes. An *H* score ranging from 0 to 100 was used to weight the similarity (edge). For two genes, G1 and G2, the *H* score was defined as follows: score(G1G2)/max(score(G1G1), score(G2G2)); the score shown here is the BLAST bit score. Gene family extraction, i.e., clustering by Hcluster_sg, used the average distance for the hierarchical clustering algorithm, which required a minimum edge weight (*H* score) of >5 and a minimum edge density (total number of edges/theoretical number of edges) of >1/3. Expansion or contraction was defined by comparing the cluster size of the ancestor to that of each of the current species using the CAFÉ program[66]. The expansions of the *PRDX1* and *OGT* genes were validated using quantitative PCR (**Supplementary Note**). tBLASTn was used to identify regions containing olfactory receptor–related sequences with at least one of the following conserved motifs: MAYDRYVAIC (TMIII), KAFSTCASH (TMVI) or PMLNPFIY (TMVII), or variants thereof with a <40% sequence difference from the conserved motifs (**Supplementary Note**).

Genome evolution. Single-copy gene families were used to construct a phylogenetic tree for *B. acutorostrata* and the other sequenced mammalian genomes. Fourfold degenerate sites were extracted from each family and concatenated to form one supergene for each species. The HKY85+gamma substitution model was selected, and PhyML v3.0 (ref. 67) was used to reconstruct the phylogenetic tree. Molecular sequence data of fourfold degenerate sites were used to estimate species divergence time using the program MCMCtree v3.0 with the approximate likelihood calculation algorithm as implemented in the PAML package[68] (version 4.5) (**Supplementary Note**).

Single amino acid polymorphisms in the minke whale and bottlenose dolphin genes were compared with those in the cow and pig genes by multiple sequence alignments using ClustalW2 (ref. 69). Protein sequences of the fin whale and finless porpoise were predicted by aligning and substituting the raw reads to the minke whale scaffolds and bottlenose dolphin scaffolds, respectively. Artifacts were removed from the alignments manually, and the filtering option required ≥1/2 coverage and ≥1/2 well-matched amino acids (the consensus string was '*', ':' or '.'). To exclude individual variation, only amino acid changes shared by all the whales tested (four minke whales and two bottlenose dolphins) were used. Significant changes in protein function ('probably or possibly damaging') were predicted using PolyPhen-2 (ref. 70).

PSGs identified on the basis of d_N/d_S ratios were predicted using branch-site likelihood ratio tests for single-copy gene families with a conservative 10% false discovery rate (FDR) criterion[10]. The minke whale was used as the foreground branch, and the cow and pig were used as the background branches for the PSGs of the minke whale. The bottlenose dolphin was used as the foreground branch for the PSGs of the bottlenose dolphin. The coding sequences of the single-copy orthologous genes were aligned using PRANK[71], and alignments shorter than 150 bp without gaps were discarded. The codeml

program in the PAML package was used to calculate the log likelihoods for the alternative model and the null model. The FDR was determined on the basis of the *q* values calculated using the *q*-value library in R[72]. All the PSGs were mapped to KEGG pathways and assigned GO terms on the basis of their *P* values, which were calculated by Fisher's exact test with a 10% FDR. The over-representation of glutathione and glutathione disulfide were validated experimentally using kidney Sp1K cells from Atlantic spotted dolphin (*S. frontalis*). Additional information regarding the methods used to identify rapidly evolving GO categories and copy number variations is provided in the **Supplementary Note**.

Demographic history. The population size histories were inferred using the PSMC model[35]. The consensus sequences of each whale were constructed and divided into nonoverlapping 100-bp bins, which were marked as homozygous or heterozygous on the basis of SNV data sets scanned using the minke whale and fin whale sequencing reads, as well as the bottlenose dolphin and finless porpoise sequencing reads mapped to the minke whale and bottlenose dolphin scaffolds, respectively. The resulting bin sequences were used as the input for PSMC estimation after removal of the sex chromosomes. Bootstrapping was performed to determine the estimation accuracy by randomly resampling 100 sequences from the original sequences. The generation times were derived from a previously published report[73].

36. Li, R. *et al.* The sequence and *de novo* assembly of the giant panda genome. *Nature* **463**, 311–317 (2010).
37. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578–579 (2011).
38. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
39. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
40. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
41. Kent, W.J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
42. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
43. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
44. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
45. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
46. Stanke, M. *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
47. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
48. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
49. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14**, 988–995 (2004).
50. Elsik, C.G. *et al.* Creating a honey bee consensus gene set. *Genome Biol.* **8**, R13 (2007).
51. Trapnell, C., Pachter, L. & Salzberg, S.L. TopHat: discovering splice junctions with RNA-seq. *Bioinformatics* **25**, 1105–1111 (2009).
52. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
53. UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **40**, D71–D75 (2012).
54. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
55. Zdobnov, E.M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
56. Bru, C. *et al.* The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* **33**, D212–D215 (2005).
57. Attwood, T.K. & Beck, M.E. PRINTS—a protein motif fingerprint database. *Protein Eng.* **7**, 841–848 (1994).
58. Attwood, T.K., Beck, M.E., Bleasby, A.J. & Parry-Smith, D.J. PRINTS—a database of protein motif fingerprints. *Nucleic Acids Res.* **22**, 3590–3596 (1994).
59. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).
60. Ponting, C.P., Schultz, J., Milpetz, F. & Bork, P. SMART: identification and annotation of domains from signaling and extracellular protein sequences. *Nucleic Acids Res.* **27**, 229–232 (1999).
61. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33**, D284–D288 (2005).
62. Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res.* **34**, D227–D230 (2006).
63. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
64. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
65. Li, H. *et al.* TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* **34**, D572–D580 (2006).
66. Hahn, M.W., Demuth, J.P. & Han, S.G. Accelerated rate of gene gain and loss in primates. *Genetics* **177**, 1941–1949 (2007).
67. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
68. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
69. Chenna, R. *et al.* Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.* **31**, 3497–3500 (2003).
70. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
71. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. USA* **102**, 10557–10562 (2005).
72. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
73. Taylor, B.L., Chivers, S.J., Larese, J. & Perrin, W.F. Generation length and percent mature estimates for IUCN assessments of cetaceans. Administrative report LJ-07-01 (Southwest Fisheries Science Center, National Marine Fisheries Service, 2007).