OPEN

Convergent evolution of the genomes of marine mammals

Andrew D Foote[1,2,16], Yue Liu[3,16], Gregg W C Thomas[4,16], Tomáš Vinař[5,16], Jessica Alföldi[6], Jixin Deng[3], Shannon Dugan[3], Cornelis E van Elk[7], Margaret E Hunter[8], Vandita Joshi[3], Ziad Khan[3], Christie Kovar[3], Sandra L Lee[3], Kerstin Lindblad-Toh[6,9], Annalaura Mancia[10,11], Rasmus Nielsen[12], Xiang Qin[3], Jiaxin Qu[3], Brian J Raney[13], Nagarjun Vijay[2], Jochen B W Wolf[2,9], Matthew W Hahn[4,14], Donna M Muzny[3], Kim C Worley[3], M Thomas P Gilbert[1,15] & Richard A Gibbs[3]

Marine mammals from different mammalian orders share several phenotypic traits adapted to the aquatic environment and therefore represent a classic example of convergent evolution. To investigate convergent evolution at the genomic level, we sequenced and performed *de novo* assembly of the genomes of three species of marine mammals (the killer whale, walrus and manatee) from three mammalian orders that share independently evolved phenotypic adaptations to a marine existence. Our comparative genomic analyses found that convergent amino acid substitutions were widespread throughout the genome and that a subset of these substitutions were in genes evolving under positive selection and putatively associated with a marine phenotype. However, we found higher levels of convergent amino acid substitutions in a control set of terrestrial sister taxa to the marine mammals. Our results suggest that, whereas convergent molecular evolution is relatively common, adaptive molecular convergence linked to phenotypic convergence is comparatively rare.

Although there are potentially several genomic routes to reach the same phenotypic outcome, it has been suggested that the genomic changes underlying convergent evolution may to some extent be reproducible and that convergent phenotypic traits may commonly arise from the same genetic changes[1–3]. Phenotypic convergence has indeed been connected to identical replacements of single amino acids within the encoded product of a protein-coding gene occurring independently in unrelated taxa[4,5]; however, such examples are rare and, to the best of our knowledge, no previous study has conducted a genome-wide scan for such convergent substitutions. Here we present high-coverage whole-genome sequences for four marine mammal species: the walrus (*Odobenus rosmarus*), the bottlenose dolphin (*Tursiops truncatus*), the killer whale (*Orcinus orca*) and the West Indian manatee (*Trichechus manatus latirostris*). These genomes provide a unique opportunity to address a key evolutionary question of what role molecular convergence has had in the evolution of shared derived phenotypic adaptations to a novel environment[6].

Mammals have evolved to inhabit the marine environment on multiple independent occasions. Cetaceans (whales, dolphins and porpoises) and sirenians (manatees and dugongs) emerged during the Eocene epoch[7–10] through diversification from the Cetartiodactyla and Afrotheria, respectively. Pinnipeds (seals, sea lions and walruses) emerged approximately 20 million years later during the Miocene from within the Carnivora[7,8]. Despite their independent evolutionary origins, pinnipeds, sirenians and cetaceans share a number of phenotypic adaptations to the pathogenic, locomotory, thermal, sensory, communication and anaerobic challenges of an aquatic existence, including limbs adapted for swimming, bone density adapted to manage buoyancy and a large total oxygen store relative to body size[6–8].

We sequenced and performed *de novo* assembly of the genomes of killer whale, manatee and walrus and increased the coverage of the previous draft genome for bottlenose dolphin by applying a whole-genome shotgun strategy using the Roche 454 and Illumina HiSeq platforms (**Supplementary Table 1**). We then predicted a set of 16,878 orthologous genes for the 4 marine mammal genomes and 6 other mammalian genomes (human, alpaca, cow, dog and elephant, with opossum as an outgroup; **Supplementary Table 2**). After filtering, we included 14,883 protein-coding orthologs for killer whale, 10,597 for dolphin, 15,396 for walrus and 14,674 for manatee.

We investigated molecular convergence among these species at two levels: first, we identified protein-coding genes evolving under positive

[1]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark. [2]Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden. [3]Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas, USA. [4]School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA. [5]Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia. [6]Broad Institute of Harvard and MIT, Cambridge, Massachusetts, USA. [7]Dolfinarium Harderwijk, Harderwijk, the Netherlands. [8]Sirenia Project, Southeast Ecological Science Center, US Geological Survey, Gainesville, Florida, USA. [9]Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [10]Marine Biomedicine and Environmental Science Center, Medical University of South Carolina, Charleston, South Carolina, USA. [11]Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy. [12]Center for Theoretical Evolutionary Genomics, University of California, Berkeley, Berkeley, California, USA. [13]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, Santa Cruz, California, USA. [14]Department of Biology, Indiana University, Bloomington, Indiana, USA. [15]Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia, Australia. [16]These authors contributed equally to this work. Correspondence should be addressed to A.D.F. (footead@gmail.com) or K.C.W. (kworley@bcm.edu).

Received 22 April 2014; accepted 29 December 2014; published online 26 January 2015; doi:10.1038/ng.3198

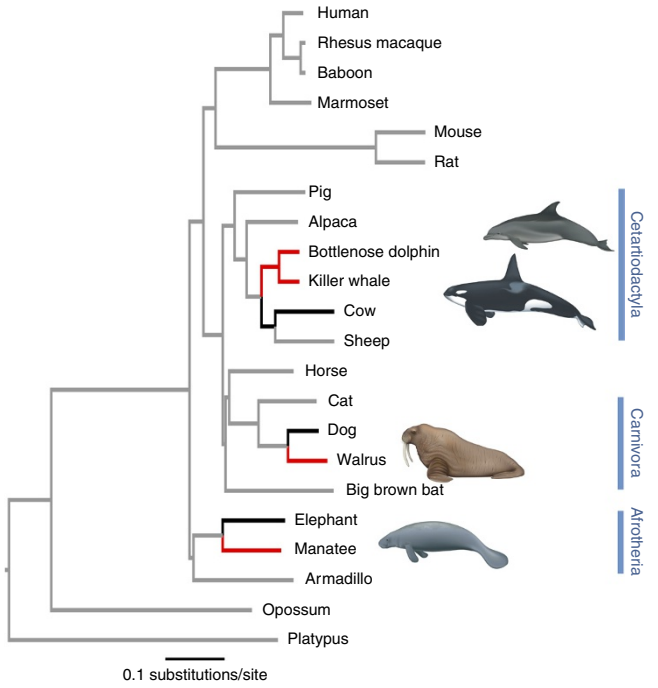

Figure 1 Phylogeny of 20 eutherian mammalian genome sequences, rooted with a marsupial outgroup. Branches representing the independent evolution of marine mammal lineages, for which tests for positive selection and parallel nonsynonymous amino acid substitutions were performed, are colored red. Branches of the control set of terrestrial taxa, for which tests for positive selection and parallel nonsynonymous amino acid substitutions were also performed, are colored black. Marine mammal illustrations are by Uko Gorter.

selection in all three orders; second, we identified convergent amino acid substitutions encoded within these genes. To identify genes evolving under positive selection, we performed a series of four different likelihood ratio tests, one on the combined marine mammal branches and one on each of the individual branches leading to manatee, walrus and the order containing dolphin and killer whale (see the branches colored red in **Fig. 1**). We identified 191 genes under positive selection across the combined marine mammal branches, 5 after conservatively correcting for multiple testing (**Supplementary Table 3**). These five genes included the glutathione metabolism pathway gene *ANPEP*. Glutathione has been experimentally demonstrated to increase antioxidant capacity in the cells of cetaceans and is thought to prevent damage by reactive oxygen species under the hypoxic conditions of long underwater dives[11] (**Supplementary Fig. 1**). There were no parallel substitutions along each of the marine mammal branches in this gene. However, we detected another glutathione metabolism pathway gene, *GCLC*, evolving under positive selection across the combined marine mammal branches (before correction for multiple testing), which encoded an identical substitution along all three branches (**Table 1**).

Such parallel nonsynonymous changes in coding genes mapping to the same amino acid site in more than one marine mammal lineage were widespread across the genome (**Fig. 2**). For example, 44 parallel nonsynonymous amino acid substitutions occurred along all 3 marine mammal lineages; these substitutions comprised 0.05% of all nonsynonymous changes. Parallel changes across any two of the marine mammal lineages occurred at an even higher rate, comprising over 1% of all changes in each combination of two marine mammal lineages (**Supplementary Table 4**). This pattern remained when we masked hypermutable CpG sites (**Supplementary Table 5**).

We found 15 of the 44 identical nonsynonymous amino acid substitutions in all 3 marine mammal lineages encoded within genes evolving under positive selection in at least one lineage; 8 of these genes were inferred to have evolved under positive selection in the test including all 3 marine mammal lineages (**Fig. 2** and **Table 1**). This finding is consistent with theoretical models that demonstrate that the probability of parallel molecular evolution increases under positive selection[12]. In all but three cases, these nonsynonymous amino acid substitutions were due to identical nucleotide changes (**Supplementary Table 6**). Only two of the nucleotide changes were at hypermutable CpG sites, and in neither case was mutation due to methylation of cytosine and subsequent deamination of resulting 5-methylcytosine to thymidine (the process that makes CpG sites prone to high mutation rates[13]). When considering the 260 non-identical amino acid substitutions affecting the same base in all 3 marine mammal lineages, we found over 25% (72) in genes evolving under positive selection in at least one marine mammal lineage (**Supplementary Table 7**).

The precise phenotypic effects of these parallel substitutions cannot currently be ascertained; however, several of the 15 genes under positive selection have known functional associations that suggest a role in the convergent phenotypic evolution of the marine mammal lineages (**Supplementary Table 8**). For example, *S100A9* and *MGP* encode calcium-binding proteins that have a role in bone formation[14,15], *SMPX* has a role in hearing and inner ear formation[16], *C7orf62* has known links to hyperthyroidism[17], *MYH7B* has a role in the formation of cardiac muscle[18] and *SERPINC1* regulates blood coagulation[19]. These genes could therefore be linked to convergent phenotypic traits such as changes in bone density (*S100A9* and *MGP*), which is high in shallow-diving species such as the manatee and walrus to overcome neutral buoyancy but low in deep-diving cetacean species that collapse their lungs to overcome neutral buoyancy. Additional functional associations of these genes with shared marine phenotypes include associations with the formation and separation of the auditory bulla

Table 1 Positively selected genes that encode parallel substitution in all three marine mammal lineages

Gene	Branch along which positive selection was detected (*P* value)	Position[a]	Convergent amino acid substitution
MYH7B	Combined marine mammals (0.0335)	1	Lys→Gln
TBC1D15	Combined marine mammals (0.0278)	15	Asn→Ser
MGP	Combined marine mammals (0.0014)	57	Leu→Ile
SMPX	Combined marine mammals (0.0315)	49	Ser→Leu
GCLC	Combined marine mammals (0.0002) and walrus (<0.0001)	220	Val→Met
SERPINC1	Combined marine mammals (0.0241), walrus (0.0400) and cetacean (0.0009)	435	Asn→Ser
M6PR	Combined marine mammals (0.0227) and cetacean (0.0242)	102	Asn→Ser
S100A9	Combined marine mammals (0.0007) and manatee (0.0051)	72	Ala→Gly
IRAK2	Cetacean (0.0091)	481	Asp→Glu
CHRM5	Cetacean (0.0449)	270	Arg→Gln
GPR97	Manatee (0.0466)	135	Ser→Arg
ESD	Manatee (0.0144)	66	Asp→Glu
SIAE	Manatee (0.0452)	415	Ile→Val
DUSP27	Walrus (0.0121)	850	Asn→Ser
C7orf62	Walrus (0.0101)	78	Ser→Asn

[a]Position of the amino acid substitution within the encoded product of the ortholog.

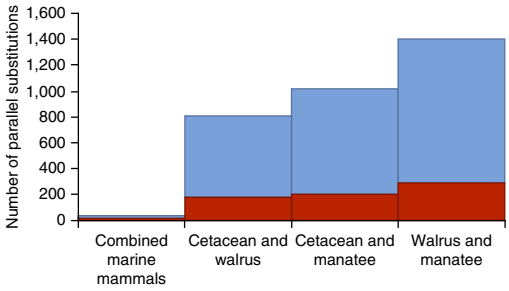

Figure 2 Genome scans for convergence. Marine mammal genomes showed a large number of parallel substitutions (blue) that occurred along the branches of at least two marine mammal lineages since they evolved from a terrestrial ancestor. Parallel substitutions that occurred in positively selected genes are shaded red.

from the skull in the inner ear (*SMPX*), unusual periodic thyroid activity (*C7orf62*), cardiovascular regulation during diving (*MYH7B*) and a low flow rate of viscous blood, particularly during diving behavior (*SERPINC1*)[7,8].

The marine mammals included in this study belong to three taxonomically distant mammalian orders; therefore, standing genetic variation and localized introgression of regions of the genome[3,20] can be ruled out as probable causes of genomic convergence. Identical *de novo* substitutions must therefore have occurred independently in each taxon during evolution from a terrestrial ancestor. Whereas most of these putatively adaptive convergent substitutions were also present in the recently published minke whale genome[11], the convergent substitutions in the *MYH7B*, *S100A9* and *GPR97* genes were not, suggesting that they were either derived in the toothed whales (Odontoceti) or lost in the baleen whales (Mysteceti) after the divergence of the Odontoceti and Mysteceti.

Surprisingly, we found an unexpectedly high level of convergence along the combined branches of the terrestrial sister taxa (cow, dog and elephant) to the marine mammals (**Supplementary Fig. 2** and **Supplementary Tables 4** and **5**), for which there is no obvious phenotypic convergence. This finding suggests that the options for both adaptive and neutral substitutions in many genes may be limited, possibly because substitutions at alternative sites have pleiotropic and deleterious effects (**Supplementary Table 8**).

Our comparison of the genomes of marine mammals has highlighted parallel molecular changes in genes evolving under positive selection and putatively associated with independently evolved, adaptive phenotypic convergence. It has been hypothesized that adaptive evolution may favor a biased subset of the available substitutions, to maximize phenotypic change[1–3], and this hypothesis may explain some of our findings of convergent molecular evolution among the marine mammals. However, we also found widespread molecular convergence among the terrestrial sister taxa, suggesting that parallel substitutions might not commonly result in phenotypic convergence. The pleiotropic and often deleterious nature of most mutations may result in the long-term survival of substitutions at a limited number of sites, leaving a signature of molecular convergence within some coding genes. The parallel substitutions in 15 positively selected genes identified in this study likely represent a small proportion of the molecular changes underlying adaptive and convergent phenotypic evolution in marine mammals. Our data therefore indicate that, although convergent phenotypic evolution can result from convergent molecular evolution, these cases are rare, and evolution more frequently makes use of different molecular pathways to reach the same phenotypic outcome.

URLs. Multi-genome alignment, ortholog set and likelihood ratio test results, http://compbio.fmph.uniba.sk/suppl/marine-mammals/; NCBI eukaryotic genome annotation pipeline, http://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/; UCSC Genome Browser, http://genome.ucsc.edu/; Baylor College of Medicine Marine Mammal Genome Project, https://www.hgsc.bcm.edu/marine-mammals; Atlas-Link, https://www.hgsc.bcm.edu/software/Atlas-Link; ATLAS GapFill, https://www.hgsc.bcm.edu/software/atlas-gapfill; Gnomon, http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml; RepeatMasker, http://www.repeatmasker.org/.

METHODS

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

Accession codes. The whole-genome shotgun sequences have been deposited in GenBank under the BioProject accessions ANOL00000000, ANOP00000000, AHIN00000000 and ABRN00000000. Sequencing data have been deposited in GenBank under BioProject PRJNA170427 corresponding to the Marine Mammal Genomes Project. Sequencing data for the Florida manatee genome have been deposited in GenBank under BioProject PRJNA189960.

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

ACKNOWLEDGMENTS

The Marine Mammals Genome Project was funded by the National Human Genome Research Institute (NHGRI), including grant U54 HG003273 (R.A.G.) for the dolphin, walrus and killer whale and grant U54 HG003067-08 for the manatee (K.L.-T. and J.A.), with additional funding from grant DNRF94 for the walrus and killer whale (M.T.P.G.) and European Union Intra-European Fellowhip (IEF) grant KWAF10 (A.D.F.). A.D.F. was supported by a Marie Curie IEF 'KWAF10' and a Lawski Foundation fellowship, and T.V. was supported by grants 1/0719/14 and 1/1085/12 from the VEGA grant agency (grant agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic). We thank the Baylor College of Medicine Human Genome Sequencing Center production teams, including those who worked on the Sanger sequencing data production teams for the dolphin (K. Abraham, S. Ali, U. Anosike, T. Attaway, D. Bandaranaike, S. Bell, B. Beltran, C. Bickham, V. Cardenas, K. Carter, I. Cavazos, M. Chandrabose, A. Chavez, J. Chu, R. Cockrell, A. Cree, M. Dao, M.L. Davila, L. Davy-Carroll, S. Denson, H. Dinh, V. Ebong, S. Fernandez, P. Fernando, N. Flagg, L. Forbes, G. Fowler, R. Gabisi, R. Garcia, T. Garner, T. Garrett, E. Hawkins, K. Hirani, M. Hogues, B. Hollins, S. Jhangiani, B. Johnson, J. Kalu, H. Kisamo, L. Lago, Y. Lai, F. Lara, T. Le, S. Lee, F. LeGall, S. Lemon, L. Lewis, L. Liu, P. London, J. Lopez, E. Martinez, C. Mercadao, M. Morgan, M. Munidasa, L. Nazareth, N. Nguyen, P. Nguyen, T. Nguyen, O. Nwaokelemeh, M. Obregon, G. Okwuonu, C. Onwere, A. Parra, S. Patil, A. Perez, Y. Perez, C. Pham, E. Primus, L.-L. Pu, M. Puazo, J. Quiroz, S. Richards, M. Ruiz, S.J. Ruiz, J. Santibanez, S. Scherer, B. Schneider, D. Simmons, I. Sisson, Z. Trejos, S. Vattathil, D. Walker, C. White, A. Williams, K. Wilson, I. Woghiren, J. Woodworth and R. Wright), the Illumina library and production teams for the walrus and killer whale (Y. Liu, S.L. Lee, S. Dugan, S. Jhangiani, D. Bandaranaike, M. Batterton, M. Bellair, C. Bess, K. Blankenburg, H. Chao, S. Denson, H. Dinh, S. Elkadiri, Q. Fu, B. Hernandez, M. Javaid, J.C. Jayaseelan, S. Lee, M. Li, X. Liu, T. Matskevitch, M. Munidasa, R. Najjar, L. Nguyen, F. Ongeri, N. Osuji, L. Perales, L.-L. Pu, M. Puazo, S. Qi, J. Quiroz, R. Raj, J. Shafer, H. Shen, N. Tabassum, L.-Y. Tang, A. Taylor, G. Weissenberger, Y.-Q. Wu, Y. Xin, Y. Zhang, Y. Zhu and X. Zou) and the submissions team (K. Wilczek-Boney, M. Batterton and D. Kalra). Large-scale computational effort was made possible by the computing cluster administered by the Center for Biomolecular Science and Engineering (CBSE) at the University of California, Santa Cruz (UCSC), funded primarily by the NHGRI, and the UPPMAX next-generation sequencing cluster and storage facility (UPPNEX), funded by the Knut and Alice Wallenberg Foundation and the Swedish National Infrastructure for Computing. Any use of trade, product or firm names is for descriptive purposes only and does not imply endorsement by the US government.

AUTHOR CONTRIBUTIONS

A.D.F. and M.T.P.G. coordinated the analyses and wrote the manuscript. K.C.W. led the sequencing consortium project. Genome assembly: Y.L., J.D., J.Q. and K.C.W.

(lead). Sequencing project managers: V.J. and S.D. Sequencing: Z.K., C.K. and D.M.M. (lead). Sequencing libraries and quality control: S.L.L. RNA sequencing analysis: X.Q. Manatee genome sequencing project: K.L.-T. and J.A. Tissue samples for dolphin: A.M. Tissue samples for walrus and killer whale: C.E.v.E. Tissue samples for manatee: M.E.H. DNA and RNA extraction (killer whale and walrus): A.D.F. Multi-genome alignment: B.J.R. Generation of ortholog set, likelihood ratio testing and GO analyses: T.V. Convergence testing: M.W.H. and G.W.C.T. Experimental design, bioinformatics and statistical support: R.N., N.V. and J.B.W.W. Additional manuscript preparation: B.J.R., M.W.H., R.A.G., N.V., T.V., J.B.W.W. and K.C.W. Principal investigators: R.A.G. and M.T.P.G.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Weinreich, D.M., Delaney, N.F., Depristo, M.A. & Hartl, D.L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312**, 111–114 (2006).
2. Tenaillon, O. *et al.* The molecular diversity of adaptive convergence. *Science* **335**, 457–461 (2012).
3. Stern, D.L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
4. Stewart, C.B., Schilling, J.W. & Wilson, A.C. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature* **330**, 401–404 (1987).
5. Wierer, M., Schrey, A.K., Kühne, R., Ulbrich, S.E. & Meyer, H.H.D. A single glycine-alanine exchange directs ligand specificity of the elephant progestin receptor. *PLoS ONE* **7**, e50350 (2012).
6. McGowen, M.R., Gatesy, J. & Wildman, D.E. Molecular evolution tracks macroevolutionary transitions in Cetacea. *Trends Ecol. Evol.* **29**, 336–346 (2014).
7. Perrin, W.F., Würsig, B. & Thewissen, J.G.M. *Encyclopedia of Marine Mammals* (Elsevier, 2008).
8. Berta, A., Sumich, J.L. & Kovacs, K.M. *Marine Mammals: Evolutionary Biology* (Academic Press, 2006).
9. 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).
10. Domning, D.P. The earliest known fully quadrupedal sirenian. *Nature* **413**, 625–627 (2001).
11. Yim, H.-S. *et al.* Minke whale genome and aquatic adaptation in cetaceans. *Nat. Genet.* **46**, 88–92 (2014).
12. Orr, H.A. The probability of parallel evolution. *Evolution* **59**, 216–220 (2005).
13. Keightley, P.D., Kryukov, G.V., Sunyaev, S., Halligan, D.L. & Gaffney, D.J. Evolutionary constraints in conserved nongenic sequences of mammals. *Genome Res.* **15**, 1373–1378 (2005).
14. Ryckman, C., Vandal, K., Rouleau, P., Talbot, M. & Tessier, P.A. Proinflammatory activities of S100: proteins S100A8, S100A9, and S100A8/A9 induce neutrophil chemotaxis and adhesion. *J. Immunol.* **170**, 3233–3242 (2003).
15. Munroe, P.B. *et al.* Mutations in the gene encoding the human matrix Gla protein cause Keutel syndrome. *Nat. Genet.* **21**, 142–144 (1999).
16. Huebner, A.K. *et al.* Nonsense mutations in *SMPX*, encoding a protein responsive to physical force, result in X-chromosomal hearing loss. *Am. J. Hum. Genet.* **88**, 621–627 (2011).
17. Eriksson, N. *et al.* Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS ONE* **7**, e34442 (2012).
18. Desjardins, P.R., Burkman, J.M., Shrager, J.B., Allmond, L.A. & Stedman, H.H. Evolutionary implications of three novel members of the human sarcomeric myosin heavy chain gene family. *Mol. Biol. Evol.* **19**, 375–393 (2002).
19. Mourey, L. *et al.* Antithrombin III: structural and functional aspects. *Biochimie* **72**, 599–608 (1990).
20. Seehausen, O. *et al.* Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192 (2014).

ONLINE METHODS

Sample collection and DNA extraction. DNA was collected from four species of marine mammals, a killer whale (*O. orca*), a bottlenose dolphin (*T. truncatus*), a Pacific walrus (*O. rosmarus divergens*) and a Florida subspecies of the West Indian manatee (*T. manatus latirostris*). The female killer whale 'Morgan' stranded on the coast of the Netherlands and was then transferred to the Harderwijk Dolfinarium. A comparison of Morgan's mitochondrial DNA sequence and learned vocal repertoire with a North Atlantic database indicated that she originated from the population of killer whales that forage primarily on the Norwegian spring-spawning stock of Atlantic herring, *Clupea harengus*[21]. A 10-ml sample of whole blood was taken and immediately stored in a PAXgene Blood DNA tube and PAXgene Blood RNA tube for DNA and RNA extraction, respectively. Additional biopsy samples from five killer whales feeding on Atlantic herring off the coast of Norway were collected and stored immediately in the preservative RNAlater. RNA was extracted and pooled from homogenized skin biopsies of the five free-ranging killer whales using the Qiagen RNeasy Mini kit and following the manufacturer's guidelines. Blood samples were similarly taken from two walruses from Harderwijk Dolfinarium: from an Alaskan male ('Igor'), with the sample immediately stored in a PAXgene Blood DNA tube for whole-genome sequencing, and from a Wrangel Island female ('Natasja'), with the sample immediately stored in a PAXgene Blood RNA tube for RNA sequencing to aid in annotation of the genome. DNA and RNA were extracted from whole blood using the PAXgene Blood DNA kit and PAXgene Blood RNA kit, respectively, and following the manufacturer's guidelines. Bottlenose dolphin tissue samples were obtained at necropsy from dolphins in the US Navy Marine Mammal Program. Spleen, liver, kidney and skin samples were from female animals, and muscle was from a male animal. Samples were used for cDNA sequencing after preparation with standard methods. Finally, blood samples were collected and DNA was extracted following standard protocols from a female Florida manatee, 'Lorelei', born in captivity and sampled at the Homosassa Springs Wildlife State Park in Homosassa, Florida, USA.

DNA and RNA sequencing and assembly. Whole-genome shotgun sequences were generated using an Illumina HiSeq platform from DNA libraries for the killer whale, walrus, manatee and bottlenose dolphin. The dolphin genome had previously undergone Sanger sequencing at 2× coverage; library and sequencing protocols have been described previously[22]. The dolphin assembly was produced by assembling the ~2.5× Sanger sequencing data with the ~3.5× Roche 454 FLX fragment data and the ~30× Illumina HiSeq data. The Sanger sequencing and 454 data were combined with the Atlas assembler, and Atlas-Link and ATLAS GapFill were then used to add the Illumina data, improve the scaffolds and fill in gaps within the scaffolds.

De novo assemblies were produced using methods similar to those applied in the Assemblathon II comparison. An initial assembly was generated using AllPath-LG with default parameters and MIN_CONTIG = 300 on all sequence data except the data for libraries with an insert size of 500 bp. The assembled scaffolds from the initial assembly were further extended using Atlas-Link on the basis of linking information provided by the libraries with insert sizes of 3 kb and 8 kb. ATLAS GapFill was then used to fill gaps within scaffolds by locally assembling the reads associated with each gap. For the killer whale and walrus, respectively, these reads were assembled into draft genomes with contig N50 sizes of 70.3 kb and 90.0 kb and scaffold N50 sizes of 12.7 Mb and 2.6 Mb (**Supplementary Table 1**). The assemblies of 2,249 Mb and 2,300 Mb covered approximately 85% and 95% of the estimated 2,373 Mb (killer whale) and 2,400 Mb (walrus) of the genomes, respectively. The improved dolphin assembly contig N50 size was 11.9 kb, and the scaffold N50 size was 115 kb. The total assembled size of the genome was 2.33 Gb (2.55 Gb with gaps) and covered ~95.3% of the genome.

Sequencing and assembly of the manatee varied slightly from the method used for the other marine mammals: the DNA from the manatee was sequenced to 90× total coverage by Illumina sequencing technology, comprising 45× coverage of libraries with a fragment size of 180 bp, 42× coverage of sheared jumping libraries with a fragment size of 3 kb, 2× coverage of sheared jumping libraries with a fragment size of 6–14 kb and 1× coverage of fosmid jumping libraries[23]. The sequence was then assembled using ALLPATHS-LG[24]. The draft assembly was 3.10 Gb in size and was composed of 2.77 Gb of sequence plus the gaps between contigs. The manatee genome assembly had a contig N50 size of 37.8 kb, a scaffold N50 size of 14.4 Mb and quality metrics comparable to those of other Illumina genome assemblies.

Annotation. The NCBI eukaryotic genome annotation pipeline was used. The first step involved repeat identification and masking with WindowMasker[25]. Second, proteins, transcripts generated from the RNA sequencing experiments and ESTs, including previously identified sequences from the study organisms or closely related organisms from RefSeq[26], were aligned to the genome assembly using BLAST. This step included a 'polishing' stage using the spice site–aware algorithm Splign[27] to improve information about splice sites and exon boundaries. Protein and transcript alignments were passed to Gnomon, which uses a hidden Markov model (HMM) tool based on Genscan[28] to extend predictions missing a start or stop codon or internal exon(s). Gnomon additionally creates *ab initio* gene predictions for regions with no evidence of alignment. The final set of annotated features comprised, in order of preference, (i) RefSeq transcripts or genomic sequences and (ii) Gnomon-predicted models. Each genome was additionally masked for repetitive elements using RepeatMasker. The proportion of repetitive elements constituting each is shown in **Supplementary Figure 3**.

Ortholog identification and alignment. The latest human (hg19), macaque (rheMac2), marmoset (calJac3), mouse (mm9), rat (rn4), alpaca (vicPac2), cow (bosTau7), dog (canFam2), elephant (loxAfr3), baboon (papAnu2) and opossum (monDom5; used as an outgroup) genome assemblies were obtained from the UCSC Genome Browser. Human-referenced whole-genome alignments were constructed from syntenic pairwise alignments with human ('syntenic nets') or reciprocal best alignments with human, depending on the quality of the assembly, using the UCSC MULTIZ alignment pipeline[29,30].

A starting gene set was composed of the human RefSeq, UCSC Known Genes[31] and VEGA[32] annotations (downloaded from UCSC on 29 July 2013). Transcripts that lacked annotated coding regions (CDSs), that had CDSs of <100 bp in length or that had CDSs whose lengths were not multiples of three were discarded. These transcripts were grouped by same-stranded CDS overlap into genes (transcript clusters). All transcripts were mapped from human to each of the other mammalian species via syntenic alignments and then subjected to a series of filters designed to minimize the impact of annotation errors, sequence quality and changes in gene structure on subsequent analyses. Briefly, each human transcript was required (i) to map to the non-human genome via a single chain of sequence alignments including ≥80% of its CDS; (ii) after mapping to a non-human species, to have ≤10% of its CDS in sequencing gaps or low-quality sequence; (iii) to have no frameshift indels, unless they were compensated for within 15 bases; and (iv) to have no in-frame stop codons and to have all splice sites conserved. To allow for genes that were mostly conserved but whose start or stop codons had shifted, incomplete transcripts with ~10% of the bases removed from the 5′ and 3′ ends of their CDSs were also considered. The final collection of ortholog sets was obtained by selecting, for each gene, the (complete or incomplete) transcript that successfully mapped to the largest number of marine mammals, with the number of other species used as a secondary criterion. In the case of a tie, the transcript with the greatest total CDS length was selected. This procedure resulted in the annotation of 16,878 genes with at least 2 non-human orthologs, with each gene present on average in ~3.3 marine species and 4.8 other species (including human).

Testing for positive selection. To find genes under positive selection, we applied four different branch-site likelihood ratio tests[33]: for the cetacean clade and branch leading to cetaceans, the walrus lineage, the manatee lineage and a single test for all branches involving the four marine mammals (foreground branches for individual tests are highlighted in **Fig. 1**). In all tests, we used the reduced parameterization introduced by Kosiol *et al.*[34]. *P* values were estimated assuming a null distribution constituting a 50:50 mixture of a χ^2 distribution and a point mass at zero, leading to conservative *P*-value estimates[35]. The Benjamini-Hochberg method[36] was used to correct for multiple testing, and a false discovery rate (FDR) cutoff of 0.1 was used. A comprehensive table of all genes in the study, together with the list of species where orthologs were found and information indicating for which tests these ortholog groups were

doi:10.1038/ng.3198

used and the resulting *P* values, and indications of whether these genes had significant FDR values after correction for multiple testing are available to download (see URLs). Genes found to be evolving under positive selection are listed in **Supplementary Tables 3** and **9–11**.

Gene Ontology (GO) categories were assigned to orthologous groups according to the human genome reference. Each gene was also assigned to all parental categories in the ontology. We used two different statistical tests to detect categories with an over-representation of positively selected genes. First, Fisher's exact test (considering all genes with $P < 0.05$ as positively selected) measured the enrichment of a particular GO category for positively selected genes (**Supplementary Table 12**). A disadvantage of this test is that its results are highly dependent on the cutoff value for positively selected genes. Second, Mann-Whitney *U* tests were used to measure shifts toward higher *P* values in a particular GO category (**Supplementary Table 13**). Thus, the Mann-Whitney *U* test does not depend on a *P*-value cutoff; however, its results may also be affected by relaxation of constraint instead of positive selection. The Holm method[37] was used to correct for multiple testing.

Testing for genomic convergence. Reconstruction of ancestral sequences was conducted for 16,833 mammalian orthologs using the Codeml program in PAMLv4.4 (ref. 38). For each of the three marine mammal groups—cetaceans, manatee and walrus—the extant sequences at each position were compared to the ancestral sequence at the node corresponding to the most recent ancestor. For the two cetaceans, this node was the one shared with cow; for walrus, this node was the one shared with dog; and, for manatee, this node was the one shared with elephant. The ancestral nodes are those at the roots of the red branches in **Figure 1**. We identified amino acid positions for which changes were inferred to have occurred and further examined those positions that changed in more than one marine mammal group. These changes could have been shared by all three groups or shared by any two of the three groups. Changes were further classified as 'parallel' if they resulted in an identical amino acid state in the present-day species and 'common' if they resulted in non-identical amino acid states in the present-day species. Common changes were hypothesized to be possible indicators of convergent evolution if adaptation to an aquatic lifestyle could be accomplished via multiple different amino acids at the same position. Genes with common and parallel changes were then compared to genes found to be under positive selection, and any overlapping genes between these two sets were inferred to have undergone convergent evolution. The positions of the parallel nonsynonymous amino acid substitutions that were found in positively selected genes are shown in **Supplementary Table 14**.

21. Foote, A.D., Kuningas, S.L. & Samarra, F.I.P. North Atlantic killer whale research; past, present and future. *J. Mar. Biol. Soc. UK* **94**, 1245–1252 (2014).
22. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
23. Williams, L.J. *et al.* Paired-end sequencing of fosmid libraries by Illumina. *Genome Res.* **22**, 2241–2249 (2012).
24. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. USA* **108**, 1513–1518 (2011).
25. Morgulis, A., Gertz, E.M., Schäffer, A.A. & Agarwala, R. WindowMasker: window-based masker for sequenced genomes. *Bioinformatics* **22**, 134–141 (2006).
26. Pruitt, K.D., Tatusova, T. & Maglott, D.R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* **35**, D61–D65 (2007).
27. Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* **3**, 20 (2008).
28. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
29. Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W. & Haussler, D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **100**, 11484–11489 (2003).
30. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14**, 708–715 (2004).
31. Hsu, F. *et al.* The UCSC Known Genes. *Bioinformatics* **22**, 1036–1046 (2006).
32. Ashurst, J.L. *et al.* The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res.* **33**, D459–D465 (2005).
33. Yang, Z. & Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917 (2002).
34. Kosiol, C. *et al.* Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**, e1000144 (2008).
35. Zhang, J., Nielsen, R. & Yang, Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479 (2005).
36. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
37. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6**, 65–70 (1979).
38. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).