Title
Convergent evolution of the genomes of marine mammals

Journal
Nature Genetics, 47(3)

ISSN
1061-4036

Authors
Foote, Andrew D
Liu, Yue
Thomas, Gregg WC
et al.

Publication Date
2015-03-01

DOI
10.1038/ng.3198

Peer reviewed

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2015 November 15.

Published in final edited form as:

Nat Genet. 2015 March ; 47(3): 272–275. doi:10.1038/ng.3198.

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], and **Richard A. Gibbs**[3]

[1]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Copenhagen, Denmark [2]Dept 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 Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [7]Dolfinarium Harderwijk, Harderwijk, Netherlands. [8]Sirenia Project, Southeast Ecological Science Center, U.S. 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, USA. [11]Department of Life Sciences and Biotechnology, University of Ferrara, Ferrara, Italy [12]Center for Theoretical Evolutionary Genomics, University of California, Berkeley, California, USA. [13]Center for Biomolecular Science and Engineering, University of California, 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

Abstract

Correspondence should be addressed to A.D. Foote (FooteAD@gmail.com) or K.C. Worley (kworley@bcm.edu).

[16]These authors contributed equally to this work.

URLs. Marine mammal genomes project, http://www.ncbi.nlm.nih.gov/bioproject/170427; Florida manatee genome, http://www.ncbi.nlm.nih.gov/bioproject/PRJNA189960; 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/; University of California, Santa Cruz genome browser, http://genome.ucsc.edu; Baylor College of Medicine marine mammal project, https://www.hgsc.bcm.edu/marine-mammals;

Accession codes. The whole-genome shotgun sequences have been deposited in GenBank under the project accessions ANOL00000000.2, ANOP00000000.1, AHIN00000000.1 and ABRN00000000.2.

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., C.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 QC: S.L.L. RNAseq 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: N.vE. 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 and likelihood ratio test & GO analyses: T.V. Convergence test: M.H. and G.T. Experimental design, bioinformatics and statistical support: R.N., N.V., J.B.W.W. Additional Manuscript Preparation: B.J.R., M.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.

Marine mammals from different mammalian orders share several phenotypic traits adapted to the aquatic environment and are therefore a classic example of convergent evolution. To investigate convergent evolution at the genomic level, we sequenced and *de novo* assembled 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 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 while convergent molecular evolution is relatively common, adaptive molecular convergence linked to phenotypic convergence is comparatively rare.

While 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 single amino acid replacements within 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 genomes of four marine mammal species: the walrus (*Odobenus rosmarus*), bottlenose dolphin (*Tursiops truncatus*), killer whale (*Orcinus orca*) and manatee (*Trichechus manatus latirostris*). These genomes provide a unique opportunity to address a key evolutionary question: the role of molecular convergence 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 walrus) emerged approximately 20 MY 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 *de novo* sequenced and assembled the genomes of killer whale, manatee and walrus, and increased the coverage of the previous draft bottlenose dolphin genome 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 four marine mammal genomes and six other mammalian genomes (human, alpaca, cow, dog, elephant and the opossum as an outgroup; Supplementary Table 2). Following filtering this resulted in the inclusion of 14,883 protein-coding orthologs for killer whale, 10,597 for the dolphin, 15,396 for the walrus and 14,674 for the manatee.

We investigated molecular convergence among these species at two levels: first, identifying protein coding genes evolving under positive selection in all three orders; second,

identifying convergent amino acid substitutions within these protein coding 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 to the order containing the dolphin and the killer whale (see branches coloured red in Fig. 1). One hundred and ninety-one genes were under positive selection across the combined marine mammal branches, five after conservatively correcting for multiple testing (Supplementary Table 3). These five included the glutathione metabolism pathway gene, *ANPEP*. Glutathione has been experimentally demonstrated to increase the antioxidant capacity in the tissue 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, another glutathione metabolism pathway gene, *GCLC*, was detected as evolving under positive selection across the combined marine mammal branches (prior to multiple testing correction) and contained an identical substitution along all three branches (Table 1).

Such parallel non-synonymous changes in coding genes at the same amino acid site in more than one marine mammal lineage were widespread across the genome (Fig. 2). For example, forty-four parallel non-synonymous amino acid substitutions occurred along all three marine mammal lineages; this comprised 0.05% of all non-synonymous 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 hypermutable CpG sites were masked (Supplementary Table 5).

Fifteen out of the forty-four identical non-synonymous amino acid substitutions found in all three marine mammal lineages were found in genes evolving under positive selection in at least one lineage; eight of these were inferred to have evolved under positive selection in the test including all three marine mammal lineages (Table 1; Fig. 2). This is consistent with theoretical models which demonstrate that the probability of parallel molecular evolution increases under positive selection[12]. In all but three cases, these non-synonymous 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 the methylation of cytosine and the subsequent deamination of 5-methylcytosine to thymidine (the process that makes CpG sites prone to high mutation rates[13]). When the 260 non-identical amino substitutions at the same base in all three marine mammal lineages were considered, over 25% (72) were found in genes evolving under positive selection in at least one marine mammal lineage (Supplementary Table 7).

The precise phenotypic effects of the parallel substitutions cannot currently be ascertained; however, several of these fifteen positively selected genes 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* are calcium binding proteins and have a role in bone formation[14,15]; *Smpx* plays 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*, *Mgp*), which is high in shallow-diving species such as the manatee and the walrus to overcome neutral buoyancy, but low in the deep-diving cetacean species that collapse their lungs to overcome neutral buoyancy. Additional functional associations of these genes with shared marine phenotypes include the formation and separation of the auditory bulla from the skull in the inner ear (*Smpx*), the unusual periodic thyroid activity (*C7orf62*), cardiovascular regulation during diving (*Myh7b*), and the low flow rate of viscous blood particularly during diving behaviour (*Serpinc1*)[7,8].

The marine mammals included in this study belong to three taxonomically distant mammalian orders; therefore genomic convergence through standing genetic variation or localised introgression of regions of the genome[3,20] can be ruled out as probable causes. Identical *de novo* substitutions must therefore have occurred independently in each taxon during their evolution from a terrestrial ancestor. While 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 they were either derived in the toothed whales (Odontoceti), or lost in the baleen whales (Mysteceti) following 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. S2, Supplementary Tables 4 and 5), along which there is no obvious phenotypic convergence. This 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 (see 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 hypothesised that adaptive evolution may favour a biased subset of the available substitutions, to maximise phenotypic change[1–3], and this 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 may 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 while 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.

Online Methods

Sample collection and DNA extraction

DNA was collected from three species of marine mammals, a killer whale (*Orcinus orca*), a bottlenose dolphin (*Tursiops truncatus*), a Pacific walrus (*Odobenus rosmarus divergens*) and a Florida subspecies of the West Indian manatee (*Trichechus 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 forages 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 5 killer whales feeding on Atlantic herring approximately off the coast of Norway, were collected and stored immediately in the preservative RNAlater. RNA was extracted and pooled from the homogenised skin biopsies of the 5 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) and immediately stored in a PAXgene Blood DNA Tube for whole genome sequencing, and from a Wrangel Island female (Natasja) and immediately stored in a PAXgene Blood RNA Tube for RNA sequencing to aid with 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 United States 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 prepared using standard methods. Lastly, blood samples were collected and DNA extracted following standard protocols from a female Florida manatee, 'Lorelei' born in captivity and sampled at the Homosassa Springs Wildlife State Park in Homosassa, FL, USA.

DNA and RNA Sequencing and Assembly

Whole genome shotgun sequences were generated using an Illumina HiSeq platform, from DNA libraries of the killer whale, walrus, manatee and bottlenose dolphin. The dolphin had previously been Sanger sequenced at 2× coverage and library and sequencing protocols have been previous described[22]. The dolphin assembly was produced by assembling the ~2.5× Sanger data with ~ 3.5× Roche 454 FLX fragment data and ~30× Illumina HiSeq data. The Sanger and 454 data were combined with the Atlas assembler and then Atlas-Link[23] and Atlas-GapFill[24] were used to add the Illumina data and improve the scaffolds and fill intra-scaffold gaps.

The *de-novo* assemblies were produced using methods similar to those used in the Assemblathon II comparison. An initial assembly was generated using AllPath-LG with default parameters and MIN_CONTIG=300 and all sequence data except the 500 bp insert data. The assembled scaffolds from the initial assembly were further extended using Atlas-

Link based upon the linking information provided from the 3 kb and 8 kb libraries. AtlasGapFill 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 cover approximately 85% and 95% of the estimated 2,373 Mb killer whale and 2,400 Mb walrus genomes respectively. The improved dolphin assembly contig N50 is 11.9 kb and the scaffold N50 is 115 kb. The total assembled size of the genome is 2.33 Gb (2.55 Gb with gaps) and covers ~95.3% of the genome.

Sequencing and assembly of the manatee varied slightly from the other marine mammals: the manatee's DNA was sequenced to 90× total coverage by Illumina sequencing technology comprising 45× coverage of 180 bp fragment libraries, 42× coverage of 3 kb sheared jumping libraries, 2× coverage of 6–14 kb sheared jumping libraries, and 1× coverage of Fosill jumping libraries (PMID: 22800726). The sequence was then assembled using ALLPATHS-LG (PMID: 21187386). The draft assembly is 3.10 Gb in size and is composed of 2.77 Gb of sequence plus gaps between contigs. The manatee genome assembly has a contig N50 size of 37.8 kb, a scaffold N50 size of 14.4 Mb, and quality metrics comparable to other Illumina genome assemblies.

Annotation

The NCBI eukaryotic genome annotation pipeline was used. The first step is repeat identification and masking using WindowMasker[25]. Second, proteins, transcripts generated from the RNA-seq 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 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 are passed to Gnomon[28], which uses a Hidden Markov model (HMM) tool based on Genscan[29] 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 alignment. The final set of annotated features comprised in order of preference, RefSeq transcripts or genomic sequences and secondly Gnomon-predicted models. Each genome was additionally masked for repetitive elements using RepeatMasker[30]. The proportion of repetitive elements constituting each is shown in Supplementary Fig. 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 as an outgroup (monDom5) genome assemblies were obtained from the University of California, Santa Cruz (UCSC) Genome Browser. Human-referenced whole-genome alignments were constructed from syntenic pairwise alignments with human (the "syntenic nets") or reciprocal-best alignments with human, depending on the quality of the assembly, using the UCSC/MULTIZ alignment pipeline[31,32].

A starting gene set was composed from of the human RefSeq, UCSC Known Genes[33], and VEGA[34] annotations (downloaded from UCSC on 29 July 2013). Transcripts that lacked annotated coding regions (CDSs), that had CDSs of <100 bp, 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 the syntenic alignments, 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 (1) to map to the non-human genome via a single chain of sequence alignments including ≥80% of its CDS; (2) after mapping to a non-human species, to have ≤10% of its CDS in sequencing gaps or low quality sequence; (3) to have no frame-shift indels, unless they were compensated for within 15 bases; (4) to have no in-frame stop codons and to have all splice sites conserved. To allow for genes that are mostly conserved, but whose start or stop codons have shifted, incomplete transcripts with ~10% of bases removed from the 5′ and 3′ ends of the CDS 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 16,878 genes with at least two non-human orthologs, averaging ~3.3 marine species and 4.8 other species (including human) per gene.

Testing for positive selection

To find genes under positive selection, we applied four different branch-site likelihood ratio tests[35]: cetacean clade and branch leading to cetaceans, walrus lineage, manatee lineage, and a single test for all branches involving the four marine mammals. In all tests we have used reduced parameterization introduced by (Kosiol *et al.*[36]). The *P*-values were estimated assuming a null-distribution that is a 50:50 mixture of a chisquared distribution and a point mass at zero, leading to conservative *P*-value estimates[37]. Benjamini and Hochberg method[38] was used to correct for multiple testing, and cutoff of false discovery rate 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 with the information indicating for which tests these ortholog groups were used and the resulting *P*-values as well as indication whether these genes were FDR significant after multiple testing correction is available to download, see **URLs**. Genes found to be evolving under positive selection are listed in Supplementary Tables 3, 9–11.

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 have used two different statistical tests to detect categories with overrepresentation of positively selected genes. First, Fisher's exact test (FET; we have considered all genes with *P*-value < 0.05 as positives) measures enrichment of a particular GO category for positives (Supplementary Table 12). A disadvantage of this test is that results will be highly dependent on the cutoff value for positively selected genes. Second, Mann-Whitney U-test (MWU) measures shifts towards higher *P*-values in a particular GO category

(Supplementary Table 13). Thus, MWU test does not depend on *P*-value cutoff, however, its results may also be affected by relaxation of constraint instead of positive selection. Holm method[39] was used to correct for multiple testing.

Testing for genomic convergence

Ancestral sequence reconstruction was conducted for 16,833 mammalian orthologs using the codeml program in PAMLv4.4[40]. For each of the three marine mammal groups—cetaceans, manatee, and walrus—at each position the extant sequences were compared to the ancestral sequence at the most recent ancestral node. For the two cetaceans, this most recent ancestral node is the one shared with cow; for the walrus, this most recent ancestral node is the one shared with dog; and for the manatee the most recent ancestral node is shared with the elephant. The ancestral nodes are indicated as those at the root of the red branches in Fig. 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 hypothesised to be possible indicators of convergent evolution if adaptation to an aquatic lifestyle can 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 parallel non-synonymous amino acid substitutions that were found in positively selected genes are shown in Supplementary Table 14.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

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 EU IEF grant KWAF10 (A.D.F.). A.D.F. was supported by a Marie Curie IEF ‘KWAF10’ and a Lawski Foundation fellowship, T.V. was supported by a grants 1/0719/14 and 1/1085/12 from the VEGA grant agency. We thank the Baylor College of Medicine Human Genome Sequencing Center production teams including those who worked on the Sanger data production teams for the dolphin (Abraham, K., Ali, S., Anosike, U., Attaway, T., Bandaranaike, D., Bell, S., Beltran, B., Bickham, C., Cardenas, V., Carter, K., Cavazos, I., Chandrabose, M., Chavez, A., Chu, J., Cockrell, R., Cree, A., Dao, M., Davila, M.L., Davy-Carroll, L., Denson, S., Dinh, H., Ebong, V., Fernandez, S., Fernando, P., Flagg, N., Forbes, L., Fowler, G., Gabisi, R., Garcia, R., Garner, T., Garrett, T., Hawkins, E., Hirani, K., Hogues, M., Hollins, B., Jhangiani, S., Johnson, B., Kalu, J., Kisamo, H., Lago, L., Lai, Y., Lara, F., Le, T., Lee, S., LeGall, F., Lemon, S., Lewis, L., Liu, L., London, P., Lopez, J., Martinez, E., Mercadao, C., Morgan, M., Munidasa, M., Nazareth, L., Nguyen, N., Nguyen, P., Nguyen, T., Nwaokelemeh, O., Obregon, M., Okwuonu, G., Onwere, C., Parra, A., Patil, S., Perez, A., Perez, Y., Pham, C., Primus, E., Pu, L.-L., Puazo, M., Quiroz, J., Richards, S., Ruiz, M., Ruiz, S.J., Santibanez, J., Scherer, S., Schneider, B., Simmons, D., Sisson, I., Trejos, Z., Vattathil, S., Walker, D., White, C., Williams, A., Wilson, K., Woghiren, I., Woodworth, J., Wright, R.), the Illumina library and production teams for the walrus and killer whale (Liu, Y., Lee, S.L., Dugan, S., Jhangiani, S., Bandaranaike, D., Batterton, M., Bellair, M., Bess, C., Blankenburg, K., Chao, H., Denson, S., Dinh, H., Elkadiri, S., Fu, Q., Hernandez, B., Javaid, M., Jayaseelan, J.C., Lee, S., Li, M., Liu, X., Matskevitch, T., Munidasa, M., Najjar, R., Nguyen, L., Ongeri, F., Osuji, N., Perales, L., Pu, L.-L., Puazo, M., Qi, S., Quiroz, J., Raj, R., Shafer, J., Shen, H., Tabassum, N., Tang, L.-Y., Taylor, A., Weissenberger, G., Wu, Y.-Q., Xin, Y., Zhang, Y., Zhu, Y., Zou, X.), the submissions team (Wilczek-Boney, K., Batterton, M., Kalra, D).

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 National Human Genome Research Institute (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 U.S. Government.

References

1. Weinreich DM, Delaney NF, Depristo MA, Hartl DL. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science. 2006; 312:111–114. [PubMed: 16601193]
2. Tenaillon O, et al. The molecular diversity of adaptive convergence. Science. 2012; 335:457–461. [PubMed: 22282810]
3. Stern DL. The genetic causes of convergent evolution. Nature Rev. Genet. 2013; 14:751–764. [PubMed: 24105273]
4. Stewart CB, Schilling JW, Wilson AC. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature. 1987; 330:401–404. [PubMed: 3120013]
5. Wierer M, Schrey AK, Kühne R, Ulbrich SE, Meyer HHD. A single glycine-alanine exchange directs ligand specificity of the elephant progestin receptor. PLOS one. 2012; 7:e50350. [PubMed: 23209719]
6. McGowan MR, Gatesy J, Wildman DE. Molecular evolution tracks macroevolutionary transitions in Cetacea. Trends Ecol. Evol. 2014; 29:336–346. [PubMed: 24794916]
7. Perrin, WF.; Würsig, B.; Thewissen, JGM. Encyclopedia of Marine Mammals. Elsevier; 2008.
8. Berta, A.; Sumich, JL.; Kovacs, KM. Marine Mammals: Evolutionary Biology. Academic Press; 2006.
9. Thewissen JG, Cooper LN, Clementz MT, Bajpai S, Tiwari BN. Whales originated from aquatic artiodactyls in the Eocene epoch of India. Nature. 2007; 450:1190–1194. [PubMed: 18097400]
10. Domning DP. The earliest known fully quadrupedal sirenian. Nature. 2001; 413:625–627. [PubMed: 11675784]
11. Yim H-S, et al. Minke whale genome and aquatic adaptation in cetaceans. Nature Genet. 2014; 46:88–92. [PubMed: 24270359]
12. Orr HA. The probability of parallel evolution. Evolution. 2005; 59:216–220. [PubMed: 15792240]
13. Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res. 2005; 15:1373–1378. [PubMed: 16204190]
14. Ryckman C, Vandal K, Rouleau P, Talbot M, Tessier PA. Proinflammatory activities of S100: proteins S100A8, S100A9, and S100A8/A9 induce neutrophil chemotaxis and adhesion. J. Immunol. 2003; 170:3233–3242. [PubMed: 12626582]
15. Munroe PB, et al. Mutations in the gene encoding the human matrix Gla protein cause Keutel syndrome. Nature Genet. 1999; 21:142–144. [PubMed: 9916809]
16. Huebner AK, et al. Nonsense mutations in SMPX, encoding a protein responsive to physical force, result in X-chromosomal hearing loss. Am. J. Hum. Genet. 2011; 88:621–627. [PubMed: 21549336]
17. Eriksson N, et al. Novel associations for hypothyroidism include known autoimmune risk loci. PLoS ONE. 2012; 7:e34442. [PubMed: 22493691]
18. Desjardins PR, Burkman JM, Shrager JB, Allmond LA, Stedman HH. Evolutionary implications of three novel members of the human sarcomeric myosin heavy chain gene family. Mol. Biol. Evol. 2002; 19:375–393. [PubMed: 11919279]
19. Mourey L, et al. Antithrombin III: structural and functional aspects. Biochimie. 1990; 72:599–608. [PubMed: 2126464]
20. Seehausen O, et al. Genomics and the origin of species. Nature Rev. Genet. 2014; 15:176–192. [PubMed: 24535286]

References

21. Foote AD, Kuningas SL, Samarra FIP. North Atlantic killer whale research; past, present and future. J. Mar. Biol. Soc. UK. 2014; 94:1245–1252.
22. Lindblad-Toh K, et al. A high-resolution map of human evolutionary constraint using 29 mammals. Nature. 2011; 478:476–482. [PubMed: 21993624]
23. Deng J, Worley KC. Atlas-Link. 2011 https://www.hgsc.bcm.edu/software/Atlas-Link.
24. Song X, Liu Y, Qu J, Gibbs RA, Worley KC. ATLAS GapFill. 2012 https://www.hgsc.bcm.edu/software/atlas-gapfill.
25. Morgulis A, Gertz EM, Schäffer AA, Agarwala R. WindowMasker: window-based masker for sequenced genomes. Bioinformatics. 2006; 2:134–141. [PubMed: 16287941]
26. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007; 35:D61–D65. [PubMed: 17130148]
27. Kapustin Y, Souvorov A, Tatusova T, Lipman D. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol. Direct. 2008; 3:20. [PubMed: 18495041]
28. Souvorov, A., et al. Gnomon - the NCBI eukaryotic gene prediction tool. National Center for Biotechnology Information. 2010. [online], http://www.ncbi.nlm.nih.gov/genome/guide/gnomon.shtml
29. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 1997; 268:78–94. [PubMed: 9149143]
30. Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996 www.repeatmasker.org.
31. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. USA. 2003; 100:11484–11489. [PubMed: 14500911]
32. Blanchette M, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004; 14:708–715. [PubMed: 15060014]
33. Hsu F, et al. The UCSC Known Genes. Bioinformatics. 2006; 22:1036–1046. [PubMed: 16500937]
34. Ashurst JL, et al. The Vertebrate Genome Annotation (Vega) database. Nucleic Acids Res. 2005; 33:459–465.
35. Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 2002; 19:908–917. [PubMed: 12032247]
36. Kosiol C, et al. Patterns of positive selection in six Mammalian genomes. PLoS Genetics. 2008; 4:e1000144. [PubMed: 18670650]
37. 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. 2005; 22:2472–2479. [PubMed: 16107592]
38. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 1995; 57:289–300.
39. Holm S. A simple sequentially rejective multiple test procedure. Scandinavian J. Stat. 1979; 6:65–70.
40. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 2007; 24:1586–1591. [PubMed: 17483113]

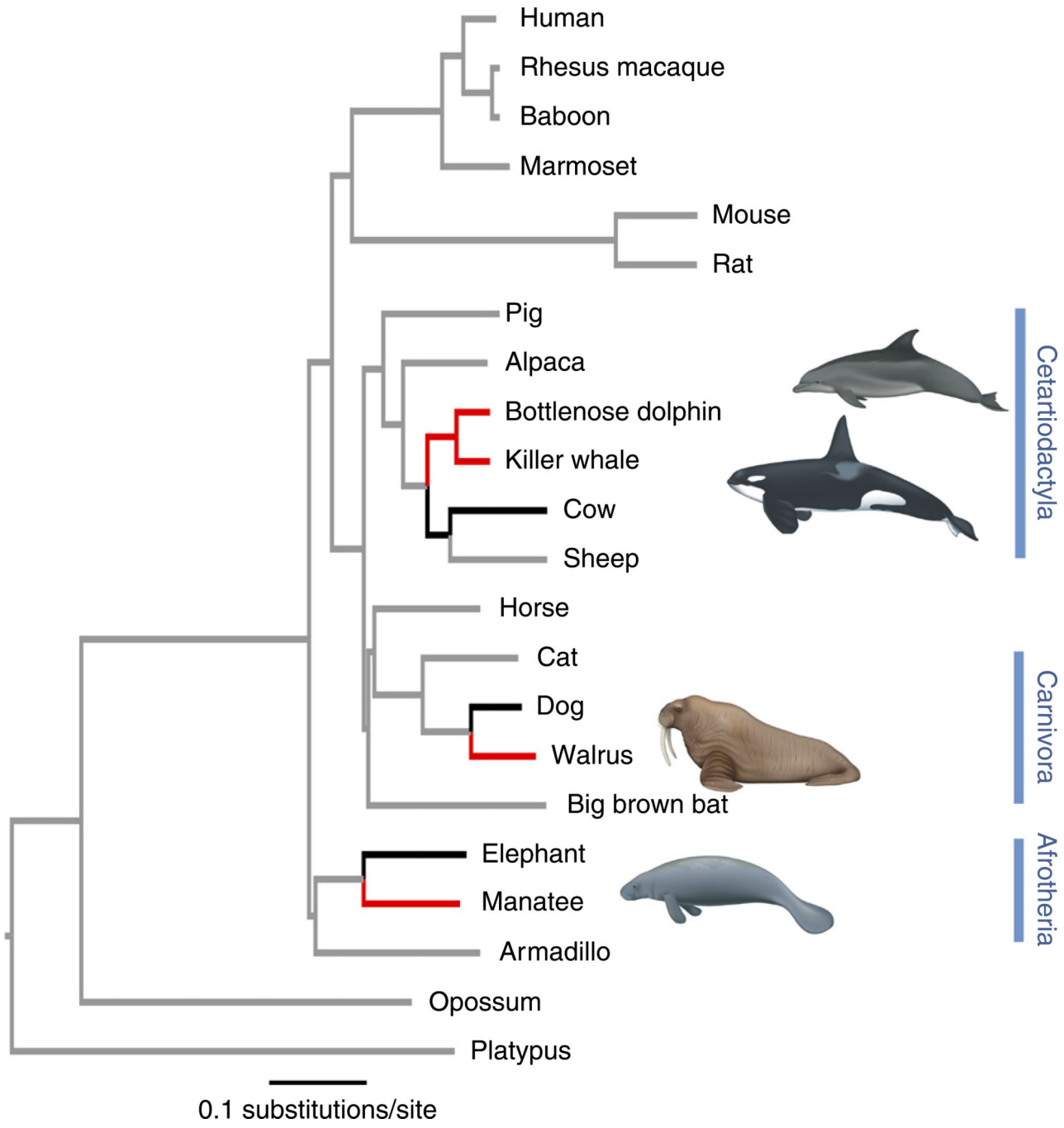

Figure 1.

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 non-synonymous amino acid substitutions were performed, are coloured red. Branches of the control set of terrestrial taxa, for which tests for positive selection and parallel non-synonymous amino acid substitutions were also performed, are coloured black. Marine mammal illustrations by Uko Gorter.

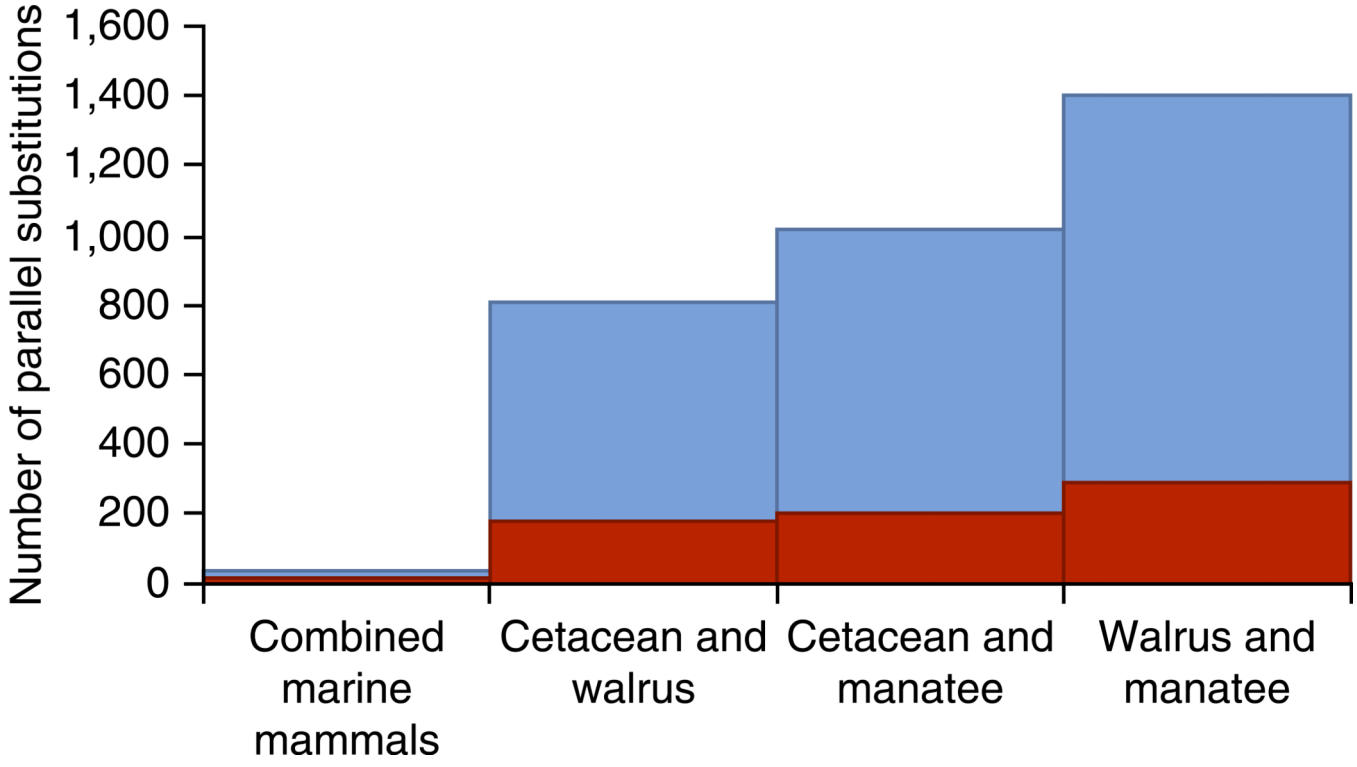

Figure 2.

Genome scans for convergence. Marine mammals genome revealed a large number of parallel substitutions (blue shaded bars) 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.

Table 1

Positively selected genes which contain parallel substitution in all three marine mammal lineages.

Gene	Branch along which positive selection was detected (*P*-value)	Position	Convergent amino acid substitution
Myh7b	combined marine mammals (0.0335)	1	K→Q
Tbc1d15	combined marine mammals (0.0278)	15	N→S
Mgp	combined marine mammals (0.0014)	57	L→I
Smpx	combined marine mammals (0.0315)	49	S→L
Gclc	combined marine mammals (0.0002) and walrus (<0.0001)	220	V→M
Serpinc1	combined marine mammals (0.0241), walrus (0.0400) and cetacean (0.0009)	435	N→S
M6pr	combined marine mammals (0.0227) and cetacean (0.0242)	102	N→S
S100a9	combined marine mammals (0.0007) and manatee (0.0051)	72	A→G
Irak2	cetacean (0.0091)	481	D→E
Chrm5	cetacean (0.0449)	270	R→Q
Gpr97	manatee (0.0466)	135	S→R
Esd	manatee (0.0144)	66	D→E
Siae	manatee (0.0452)	415	I→V
Dusp27	walrus (0.0121)	850	N→S
C7orf62	walrus (0.0101)	78	S→N