The novel evolution of the sperm whale genome

Wesley C. Warren¹, Lukas Kuderna², Alana Alexander³, Julian Catchen⁴, José G. Pérez-Silva⁵, Carlos López-Ótin⁵, Víctor Quesada⁵, Patrick Minx¹, Chad Tomlinson¹, Michael J. Montague⁶, Fabiana H.G. Farias¹, Ronald B. Walter⁷, Tomas Marques-Bonet², Travis Glenn⁸, Troy J. Kieran⁸, Sandra S. Wise⁹, John Pierce Wise, Jr.⁹, Robert M. Waterhouse¹⁰, John Pierce Wise, Sr.⁹

¹McDonnell Genome Institute, Washington University, 4444 Forest Park Ave., St Louis, MO 63108, USA.
²Institute of Evolutionary Biology (UPF-CSIC), PRBB, 08003 Barcelona, Spain.
³Biodiversity Institute, University of Kansas, 1345 Jayhawk Blvd., Lawrence, KS, 66047, USA.
⁴Department of Animal Biology, University of Illinois, Urbana, IL, 61801, USA.
⁵Departamento de Bioquímica y Biología Molecular, Facultad de Medicina, Instituto Universitario de Oncología, Universidad de Oviedo, 33006-Oviedo, Spain.
⁶Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.
⁷Department of Chemistry and Biochemistry, Texas State University, 419 Centennial Hall, 601 University Drive, San Marcos, TX, 78666, USA.
⁸Department of Environmental Health Science, University of Georgia, Environmental Health Science Bldg., Athens, GA, 30602, USA.
⁹Wise Laboratory of Environmental and Genetic Toxicology, Department of Pharmacology and Toxicology, School of Medicine, University of Louisville, Louisville, KY 40292, USA.
¹⁰Department of Ecology and Evolution, University of Lausanne, JO, CH-1015 Lausanne, Switzerland.

© The Author(s) 2017. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com.
The sperm whale, made famous by *Moby Dick*, is one of the most fascinating of all ocean-dwelling species given their unique life history, novel physiological adaptations to hunting squid at extreme ocean depths, and their position as one of the earliest branching toothed whales (Odontoceti). We assembled the sperm whale (*Physeter macrocephalus*) genome and resequenced individuals from multiple ocean basins to identify new candidate genes for adaptation to an aquatic environment and infer demographic history. Genes crucial for skin integrity appeared to be particularly important in both the sperm whale and other cetaceans. We also find sperm whales experienced a steep population decline during the early Pleistocene epoch. These genomic data add new comparative insight into the evolution of whales.

**Introduction**

The sperm whale, made famous by *Moby Dick*, makes some of the deepest and longest dives of any marine mammal: >73 min long and up to 2,035 m deep (Watkins 1993; Watwood et al. 2006) to feed on squid, including the infamous giant and colossal squids (Best 1979; Whitehead 2003). Previous comparative genomic analyses of cetaceans indicated genic adaptation to a marine existence (Foote et al. 2015; Yim et al. 2014), including convergent pathways of metabolism regulation for deep diving (Foote et al. 2015). However, to date, the sperm whale – one of the deepest diving and earliest branching toothed whales (Odontoceti) (Whitehead 2003) – has been excluded from these comparisons. We sequenced and assembled multiple sperm whale genomes to explore genic adaptation. Given the important and broad physiological roles played by proteases, our explorations mostly focused on examining protease loss-of-function events important in sperm whale, and cetacean, evolution. We also sought to discover which genes showed signs of positive selection shared with other cetaceans or unique to sperm whale. Finally, as previous analyses suggested that sperm whales experienced a global expansion <80,000 years ago (Alexander et al. 2016), we examine the estimated historical effective population size using samples from throughout the sperm whale’s range.

**Materials and Methods**

We sequenced a Gulf of Mexico female sperm whale (GMX) to high coverage (72x) using short-insert and mate-pair libraries of 100 bp length (detailed in the Supplementary Materials) on an Illumina HiSeq2000. We assembled the draft genome of all sequences with ALLPATHS
(Gnerre et al. 2011) using default parameter settings, subjecting assembly input reads to quality control as detailed in the ALLPATHS documentation (Gnerre et al. 2011). We obtained RNAseq data from skin biopsies of a different GMX sperm whale to aid gene annotation as described in the Supplementary Materials. Gene annotation was performed according to the NCBI gene annotation pipeline as described here: http://www.ncbi.nlm.nih.gov/books/NBK169439/. After aligning genes from the sperm whale with other taxa (detailed in Supplementary Materials) to establish 1:1 gene orthology, positive selection was detected using PAML4.0 (Yang 2007) and impact on protein structure tested with Provean (Choi and Chan 2015). Canonical pathway enrichment of gene clusters under positive selection was established as detailed in the Supplementary Materials. Protease genes were manually annotated and validated for loss/duplication events using BATI (http://degradome.uniovi.es/downloads.html). Four additional sperm whale individuals (Supplementary Table 1) were re-sequenced to moderate depth (21-28×) and reads were mapped to the draft genome as described in the Supplementary Materials. We calculated heterozygosity on a per-individual basis using VCFtools (Danecek et al. 2011). Effective population size was reconstructed with PSMC (Li and Durbin 2011) using the parameters specified in the Supplementary Materials.

**Results and Discussion**

Our sperm whale total assembled sequence was similar in size to other assembled cetacean genomes (Supplementary Table 2). Using our Gulf of Mexico individual (GMX) reference assembly (Genbank assembly accession GCA_000472045.1) we inferred 18,686 protein-coding genes – second only to the baiji (Lipotes vexillifer) among sequenced cetaceans at 18,906 genes. Using a core eukaryotic mapping method (Simao et al. 2015) we also demonstrated >94.7% of conserved genes were complete in our assembly (Supplementary Table 3). Of the 18,686 protein-coding genes, 12,717 had single-copy orthologs in both human and other cetartiodactyls (Supplementary Table 4; additional methods/results can be found in the Supplementary Materials). A total of 45 genes found across 8 taxa were identified as being under positive selection in the sperm whale lineage; these genes also passed our stringent functional impact tests (default cutoff < -2.5) using Provean (Choi and Chan 2015) (Supplemental File 1). Several significant pathways emerged from enrichment analyses, which included genes associated with blood-circulation and skin stress responses (Table 1).
Cetaceans, including the sperm whale, exhibit molting or skin sloughing (Amos 1992), potentially as an adaptive response to fouling by barnacles and other organisms. However, sperm whales face the additional challenge of maintaining skin integrity and blood homeostasis at high water pressures during deep foraging dives.

To complement the analysis of genes under positive selection, we manually annotated the complete set of proteases (i.e. degradome) of the sperm whale. This independent analysis identified several proteases involved in skin function and blood homeostasis that showed loss-of-function (LoF) events along the lineage leading to sperm whales (Figure 1; additional methods/results in Supplementary Materials). We also detected LoF in proteases involved in inflammation, immunity and metabolism within cetaceans, and specifically within the sperm whale. A loss of several proteases in cetaceans suggests a trend towards a milder inflammatory response relevant to Peto’s paradox: a theory postulated to explain the lower relative incidence of cancer in large mammals (Caulin and Maley 2011). In addition, MMP7 – which promotes metastasis when expressed at high levels (Li et al. 2014; Koskensalo et al. 2011) – contains a premature stop codon in sperm whales, a putative sperm whale-specific mechanism to reduce cancer incidence. We also found that CASP12 and PRSS33 were independently lost in cetaceans and some hominoids, suggesting a case of convergent evolution of the immune system in very different environments. Several proteases involved in digestion (CPA2, CPA3, CPO) were also lost in cetaceans (Figure 1). In some cases these losses were independent, suggesting convergent evolution driven by trophic level. As expected, odontocetes retain functional orthologs of proteases involved in dentition (KLK4, MMP20), which were lost in mysticetes, who use baleen – not teeth – to filter food (Keane et al. 2015).

To better understand patterns of genetic diversity among sperm whales from different ocean basins, we carried out medium-coverage resequencing of individuals from the Pacific Ocean and Indian Ocean. Average genome-wide heterozygosity per base, corrected for callable sequence space, was 0.0011. This value is low in comparison with the fin whale (0.0015) and bottlenose dolphin (0.0014) (Yim et al. 2014), suggesting the sperm whale has a smaller effective population size ($N_e$). A pairwise sequentially Markovian coalescent (PSMC) analysis (Li and Durbin 2011) indicated a rapid decline in $N_e$ during the transition from the Pliocene to Pleistocene epoch, inferred consistently regardless of the ocean origin of samples (Figure 2A). The increase in upwelling associated with the Pliocene and/or cycles of glaciation within the
Pleistocene have been implicated in the evolution of gigantism in mysticetes (Slater et al. 2017), as well as the diversification of marine dolphins (do Amaral 2016). This suggests that changes in ocean dynamics during this time period have had a strong impact on cetaceans in general, and we suggest are also the likely cause of the inferred sperm whale population decline. The GMX sample had significantly lower heterozygosity than Pacific and Indian Ocean samples (Figure 2B, Supplementary Table 1). Future sequencing will clarify whether lower diversity is restricted to GMX, or characteristic of the entire Atlantic. However, the isolation of GMX due to high levels of female philopatry (inferred from differentiation of the maternally-inherited mitochondrial DNA (Engelhaupt et al. 2009; Alexander et al. 2016)), and the limited census size (763 sperm whales in 2009 (Waring 2013)), suggest that GMX could be subjected to greater levels of genetic drift associated with a small and maternally-isolated population. The ability of the sperm whale to respond to future selective pressures, including climate change, in the face of such reduced genetic diversity should be a focus of ongoing study.

Overall, our results suggest positive selection has differentially affected localized portions of the sperm whale genome. In particular, the complex pattern of convergent gene evolution involving skin-related genes suggests they have played an important role in aquatic adaptation, possibly influenced by the somewhat contradictory requirements of heat insulation, buoyancy and deep diving. In comparison to the localized effects of selection on the genome, we infer that the sperm whale experienced a rapid population decline, potentially in response to glaciation, which had a broad effect on genome-wide diversity. Given the apparent influence of past climate change, monitoring the on-going response of sperm whales to anthropogenically mediated climate change will be paramount.
Table 1. Genes under positive selection enriched by pathway, phenotype or protein interactions.

<table>
<thead>
<tr>
<th>Pathway</th>
<th>Source</th>
<th>Genes</th>
<th>Ratio of enrichment</th>
<th>Adjusted P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focal adhesion</td>
<td>Wiki</td>
<td>CHAD, COL1A2, THBS2, TNC, FLT1</td>
<td>6.9</td>
<td>0.015</td>
</tr>
<tr>
<td>Focal adhesion</td>
<td>KEGG</td>
<td>CHAD, COL1A2, PARVG, THBS2, TNC, FLT1</td>
<td>7.5</td>
<td>0.0076</td>
</tr>
<tr>
<td>Pemphigus</td>
<td>Disease</td>
<td>PPL, EVPL, DSP, DSG3</td>
<td>38.8</td>
<td>0.0006</td>
</tr>
<tr>
<td>Calcium signaling</td>
<td>KEGG</td>
<td>PTK2B, ADCY3, RYR1, NOS2, P2RX3, PHKB</td>
<td>7.6</td>
<td>0.0076</td>
</tr>
<tr>
<td>Blood circulation</td>
<td>GO</td>
<td>ALOX5, CHD7, WNK1, EPAS1, CX3CL1, PPP1R13L, COL1A2, AZU1, DSP, GUCY1A3, MYBPC3, TBC1D8</td>
<td>3.4</td>
<td>0.036</td>
</tr>
<tr>
<td>Cornified envelope</td>
<td>Reactome</td>
<td>DSP, TGM1, KRT4, PKP1, DSG3, PPL, EVPL</td>
<td>NA</td>
<td>(^1\text{FDR} 3.65\times10^{-11})</td>
</tr>
</tbody>
</table>

\(^1\) The false discovery rate (FDR) calculated within the Reactome software (Croft 2014) is the probability corrected for multiple comparisons. Adjusted P values are not provided.
Figure 1. Cetacean-specific losses of protease genes. Proteases that have undergone loss-of-function in sperm whales, specifically, are shown to the left of the phylogeny while those that are inferred to be convergent, or inferred to have occurred in ancestral lineages, are mapped on to the phylogeny. Each event is depicted along the branch where loss events have been inferred to occur. Genes expected to impact skin function are colored blue; immune system: purple; blood homeostasis: red; digestion: orange, and those showing convergent loss-of-function as underlined bold. The unique duplication of sperm whale CASP3 is shown above the phylogeny and marked by an asterisk.
Figure 2. Estimated effective population size history and heterozygosity of sperm whales from different ocean basins. Samples are color coded by the key, with blue/green = Pacific, orange/yellow = Indian Ocean, and dark orange = Gulf of Mexico, Atlantic. A) PSMC reconstruction of effective population size through time by sample (excluding SEY420021031-063, see Supplementary Materials), dashed lines represent the estimated start dates for each epoch; B) Genome wide distribution of heterozygosity for each sample, by contig/scaffold. The Gulf of Mexico sample – characterized by low heterozygosity – is marked by an asterisk where it has the largest number of contigs in a category. The insert emphasizes that this sample has the largest number of contigs with low heterozygosity (< 0.0005).
Authors contributions

SSW, JWJ, JWS isolated genomic DNA and sexed all samples. LK, TMB performed all population history analyses. JGP, CLO, VQ performed all protease analyses. CT, PM completed genome assembly and curation. JC, TJK, TG analysed population sequences. MJM, FHF analysed gene selection. RMW completed all gene orthology analyses. WCW, AA wrote the paper and reviewed all analyses. All authors have read and approved the manuscript.

Acknowledgments

This work was supported by the National Institute of Health grant 2R24OD011198-04A1 (W.C. Warren, PI), National Institute of Environmental Health Sciences [ES016893 (J. Wise Sr., PI)]; Army Research Office [W911NF-09-1-0296 (J. Wise Sr., PI)]; Swiss National Science Foundation grant PP00P3_170664 (R.M. Waterhouse); Ministerio de Economia and Competitividad (Spain) and European Union (ERC-Advanced Grant DeAge) (C. López-Otín, PI); the Campbell Foundation; the Ocean Foundation; Ocean Alliance; and the many individual and anonymous Wise Laboratory donors. The authors declare no competing financial interests. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Environmental Health Sciences, the National Institutes of Health, the Army Research Office or the Department of Defense. Work was conducted under National Marine Fisheries Service permit #1008-1637-03 (J. Wise Sr., PI) and permit #751-1614 (Iain Kerr, PI). We thank Kyung Kim for computational support. We would like to thank Catherine Wise, James Wise, Chris Gianios and all the Wise Laboratory/Odyssey science team volunteers for their help with technical support, whale spotting and sample collection. We thank Iain Kerr, Roger Payne, Bob Wallace, Derek Walker, and all of the Odyssey boat crew for their help with sample collection and logistics. We thank C. Scott Baker for comments on this manuscript. Finally, we thank all the volunteers who supported us in this project.

References


