Genome-Wide Scans for Candidate Genes Involved to the Aquatic Adaptation of Dolphins

Yan-Bo Sun\textsuperscript{1,3,#}, Wei-Ping Zhou\textsuperscript{1,2,3,#} He-Qun Liu\textsuperscript{1,3,6} David M. Irwin\textsuperscript{1,4,5} Yong-Yi Shen\textsuperscript{1,7,*}, and Ya-Ping Zhang\textsuperscript{1,3,*}

\textsuperscript{1} State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, the Chinese Academy of Sciences, Kunming 650223, China;
\textsuperscript{2} Department of Molecular and Cell Biology, School of Life Sciences, University of Science and Technology of China, Hefei, Anhui 230026, China;
\textsuperscript{3} Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming 650091, China;
\textsuperscript{4} Department of Laboratory Medicine and Pathobiology, University of Toronto, Ontario, Canada;
\textsuperscript{5} Banting and Best Diabetes Centre, University of Toronto, Ontario, Canada;
\textsuperscript{6} Graduate School of the Chinese Academy of Sciences, Beijing, China;
\textsuperscript{7} School of Life Sciences, Xiamen University, Xiamen 361005, China;
\# These authors contribute equally.

* Corresponding author: Yong-Yi Shen (shen_yongyi@yahoo.com.cn)

or Ya-Ping Zhang (zhangyp@mail.kiz.ac.cn)

Keywords: \textit{Tursiops truncates}, aquatic adaptation, positive selection, branch-site model
Abstract

Since their divergence from the terrestrial artiodactyls, cetaceans have fully adapted to an aquatic lifestyle, which represents one of the most dramatic transformations in mammalian evolutionary history. Numerous morphological and physiological characters of cetaceans have been acquired in response to this drastic habitat transition, such as thickened blubber, echolocation, and ability to hold their breath for a long period of time. However, knowledge about the molecular basis underlying these adaptations is still limited. The sequence of the genome of *Tursiops truncates* provides an opportunity for a comparative genomic analyses to examine the molecular adaptation of this species. Here, we constructed 11,838 high-quality orthologous gene alignments culled from the dolphin and four other terrestrial mammalian genomes and screened for positive selection occurring in the dolphin lineage. A total of 3.1% (368) of the genes were identified as having undergone positive selection by the branch-site model. Functional characterization of these genes showed that they are significantly enriched in the categories of lipid transport and localization, ATPase activity, sense perception of sound, and muscle contraction, areas that are potentially related to cetacean adaptations. In contrast, we did not find a similar pattern in the cow, a closely related species. We re-sequenced some of the positively selected sites (PSSs), within the positively selected genes (PSGs), and showed that most of our identified PSSs (50/52) could be replicated. The results from this study should have important implications for our understanding of cetacean evolution and their adaptations to the aquatic environment.

Introduction

Cetaceans diverged from artiodactyls approximately 50 million years ago (Meredith et al. 2011) and their habitat transition, from land to an aquatic environment, represents one of the most dramatic transformations in mammalian evolutionary history. These adaptation inevitably posed challenges for the ancient cetaceans, which had originally been adapted for terrestrial life, with locomotion (navigation) and detection of prey being major ones. For locomotion, they needed to confront the
considerable obstacle provided by water, whose density is much higher than air. To overcome drag, cetaceans have evolved some extreme changes in morphology and physiology, including a streamlined form and a modified skeletal system (Fish, Beneski, Ketten 2007; Reidenberg 2007). In addition, most cetaceans possess a thick layer of blubber, which increases their buoyancy (Struntz et al. 2004). For foraging in water, these mammals constantly need to hunt at night or in deep water, therefore, it is vital for them to possess superior capabilities of long-time diving and locating prey. It is striking that some cetacean species have acquired an ability to echolocate that has enabled them to use sound to locate prey or escape obstacles when navigating (Cranford, Amundin, Norris 1996). Moreover, cetaceans have elevated levels of myoglobin in their skeletal muscles (Noren et al. 2001; Wright, Davis 2006), which vastly increases their ability to retain oxygen, allowing for longer time between breaths. They also utilize glycolysis metabolism to compensate for insufficient levels of oxygen (Butler, Jones 1997), which potentially supports the energy supply for their long dives.

Considering the significant phenotypic modifications in cetaceans, it should be expected that these modifications were shaped by natural selection, and conferred a selective advantage, as they adapted to the new aquatic environment. What are the underlying molecular mechanisms for these innovations? Positive Darwinian selection is one of the major driving forces for adaptive evolution and species diversification, which had been widely investigated in many species (Kosiol et al. 2008; Lefebure, Stanhope 2009; Shen et al. 2010; Oliver et al. 2011; Wissler et al. 2011; McGowen, Grossman, Wildman 2012). A few studies have focused on the adaptive evolution of marine mammals (McClellan et al. 2005; Wang et al. 2009), however, as the complete genomes of marine mammals were not available at that time, the datasets analyzed in these previous studies were limited to only a few genes (eg. cytB and HoxD) (McClellan et al. 2005; Wang et al. 2009). Whole-genome wide identifications of positively selected genes (PSGs) along the marine mammal lineage should greatly help us understand the genetic bases underlying adaptive evolution in marine mammals. The genome sequence of the bottlenose dolphin (Tursiops truncates)
provides an opportunity to conduct this analysis.

A series of evolutionary models for testing positive selection have been developed in the past decade, including the branch model, the site model, and the branch-site model (Yang 1998; Yang et al. 2000; Yang, Nielsen 2002; Yang, Wong, Nielsen 2005; Zhang, Nielsen, Yang 2005). In the first two models, positive selection is inferred only if the $dN/dS$ average over all sites or all branches is significantly greater than 1. Positive selection, however, often operates episodically on only a small number of sites on a few lineages (Yang, Nielsen 2002), limiting the power of detecting positive selection by the branch and site models. The branch-site model, a more powerful model, was developed to address this issue (Yang, Nielsen 2002; Zhang, Nielsen, Yang 2005) and has been widely used in screens for positive selection [cf. (Bakewell, Shi, Zhang 2007; Kosiol et al. 2008; Studer et al. 2008; Shen et al. 2010)].

Here, we constructed whole-genome ortholog gene sets among five mammalian species, including dolphin (*Tursiops truncates*), cow (*Bos Taurus*), dog (*Canis familiaris*), panda (*Ailuropoda melanoleuca*), and human (*Homo sapiens*), and identified positively selected genes (PSGs) along the dolphin lineage with the improved branch-site model to build a database of genes that might be correlated with aquatic adaptation in the dolphin. As the current release of the dolphin genome has only 2.59× coverage, there are limitations for comparative genomic analyses, especially the detection of positive selection. Sequencing errors, problems with annotation, alternative splicing, amino acid repeats, and frameshift mutations could generate a higher rate of false-positive with the branch-site model (Mallick et al. 2009; Schneider et al. 2009; Markova-Raina, Petrov 2011), therefore, generating accurate alignments is an essential step in the inference of positive selection. The Prank software (Loytynoja, Goldman 2005; Loytynoja, Goldman 2008) was recently reported as being able to generate much more accurate alignments than other traditional aligners (Fletcher, Yang 2010; Markova-Raina, Petrov 2011), thus we used this algorithm to align all the genes used in this study. Moreover, we re-sequenced some of the candidate PSS regions to confirm their reliability. We show that most
(50/52) of our identified PSSs are reliable. Through a functional clustering analysis of the dolphin PSGs, we found that they are enriched for categories such as lipid transport and localization, ATPase activity, perception of sound, and muscle contraction clusters.

**Material and Methods**

Coding region sequences of individual genes from the genomes of the dolphin and other species were downloaded from Ensembl (version 66, March 2012) using the BioMart tool (Vilella et al. 2009). The species used here for comparison with dolphin include cow (*Bos Taurus*, UMD3.1), dog (*Canis lupus familiaris*, CanFam_2.0), panda (*Ailuropoda melanoleuca*, ailMel1), and human (*Homo sapiens*, GRCh37.p6). A phylogenetic tree of these species is shown in Figure 1, which is derived from Murphy et al. (Murphy et al. 2007). To predict homologs among the five genomes, we used the Ensembl inferences (Vilella et al. 2009). For each pair of these genomes, only those that Ensembl annotated as one2one orthologous genes were retrieved and analyzed in the following step. If a gene had multiple transcripts, then the longest one was chosen. After these treatments, we obtained 12,057 gene sets.

The Prank program (Loytynoja, Goldman 2005; Loytynoja, Goldman 2008) was used to align all of the gene sets. Since Prank performs much better at the codon level than at the amino-acid level (Fletcher, Yang 2010) for protein-coding sequences, all of the genes were thus directly aligned at the codon level with the option “-codon”.

After the alignments were generated, we performed a trimming treatment to remove potentially unreliable regions using the Gblocks program (Castresana 2000). The parameters used were the default settings with the sequence type being codon (“-t=c”). In addition, to reduce the effect of uncertain bases on the inference of positive selection, we deleted all positions that had gaps (“—”) and “N” from the alignments. After the trimming process, if the remaining alignment was shorter than 120bp (40 codons), then the entire alignment was discarded.

In addition to alignment uncertainty, saturation at silent sites (*d**S**) may also bias the inference of positive selection (Smith, Smith 1996). To identify saturation, for
each gene the third codon positions were extracted and branch lengths on the species
tree were estimated using the GTR model with PAML (Yang 2007). Branch lengths
were used as a proxy for saturation, and genes were removed from the analysis if one
or more branches had a length ≥ 1. Our final dataset contained 11,838 genes.

For each of the remaining genes, a branch-site evolutionary analysis for positive
selection was conducted using codeml from the PAML package (Yang 2007). In this
study, the improved branch-site model (Yang, Nielsen 2002; Zhang, Nielsen, Yang
2005) was used. This model requires that the branches of the tree be divided in priori
into foreground and background lineages. A likelihood ratio test (LRT) compares a
model with positive selection on the foreground branch to a null model where no
positive selection occurred on the foreground branch and calculates the statistic
(2Δln) to obtain a P-value. In this study, genes were inferred to be PSGs only if the
P-value was less than 0.01. This model can also infer positively selected sites (PSS)
based on an empirical Bayes analysis (Yang, Wong, Nielsen 2005). In this study, PSS
were inferred only if their posterior probability was greater than 95%.

After PSGs were detected, we used the DAVID Functional Annotation tool
(Huang da, Sherman, Lempicki 2009) to investigate their enrichment of Gene
Ontology (GO) terms. During this analysis, the human ortholog of the PSG was as the
input against a background of human genes. Within each annotation cluster, DAVID
lists the GO terms that are significantly enriched. In this study we used the approach
of McGowen et al. (McGowen, Grossman, Wildman 2012), where only terms with an
enrichment score > 1.3 were considered meaningful.

To confirm our identified positively selected sites, we randomly selected 48
PSGs for whom the presence of PSS had been detected, and designed PCR primers
using Primer3 (Rozen, Skaletsky 2000) to directly amplify and sequence these regions
using PCR and an Applied Biosystems 3730 DNA Analyzer, respectively. DNA for
this study was extracted from the same species of dolphin (Tursiops truncatus).
Information on these primers is available in Supplementary Table 1 and all the
segments sequenced in this study were deposited into GenBank with accession
numbers from JX856347 to JX856394.
Results

Our analysis began with 12,057 genes that had single copy orthologs in the dolphin, cow, dog, panda, and human genomes. Each of these genes was annotated by Ensembl (version 66) as being one2one orthologous between each pair of species. After our alignment treatments, 219 genes were eliminated because either their final lengths were shorter than 120bp or they did not pass the synonymous substitution saturation test (see methods). Finally, 11,838 genes were tested for positive selection in this study.

First, to determine the overall difference in selective constraints between dolphin and other species, each aligned gene was evaluated for their substitution rates including \(dN\), \(dS\) and \(dN/dS\), under the species tree (Figure 1). The free-ratio model (M1) in PAML (Yang 2007), which allows a separate \(\omega\) for each branch, was used. We found that 97% of the genes had a \(dN/dS\) ratio smaller than 1 on the dolphin lineage, providing support for the overall presence of purifying selection acting at the molecular level over all time. We then compared the mean \(dN/dS\) between dolphin and its closely related species. When \(dS\) is approximately, or equal to, 0 along a branch, then it always generates a very high \(dN/dS\) value, hence, for this comparison, genes with \(dS < 0.0005\) were not included. The mean \(dN/dS\) along the dolphin lineage was 0.2373, significantly larger than 0.1435 on the cow lineage (\(P < 0.001\), Mann-Whitney Test; Figure 2), suggesting that genes evolved faster in the dolphin after their split from artiodactyls. Indeed, by examining the \(dN/dS\) for each gene on the dolphin and cow lineages, we found 8,422 genes that have higher \(dN/dS\) in the dolphin and only 2,940 genes where it is higher in the cow. In addition, the mean rate of synonymous substitutions along the dolphin lineage was found to be much smaller than those measured for of the other species used in this study (Figure 2), which agrees with a previously reported slowdown in the molecular rate in dolphin (McGowen, Grossman, Wildman 2012). The full list of genes, with their substitution traits, is available in Supplementary Table 2.

Next, we utilized the codeml program in the PAML package (Yang 2007), with
the improved branch-site model ("test2") (Zhang, Nielsen, Yang 2005), to detect signals of positive selection on each alignment. These screens identified 368 genes (3.1% of the total) that show significant evidence of positive selection (\(P < 0.01\)) in the dolphin lineage (Supplementary Table 2). The PSGs were applied to the DAVID Functional Annotation tool (Huang da, Sherman, Lempicki 2009) to investigate their functional enrichments. Interestingly, we found a number of functional categories that might be correlated with the dolphin-specific traits that were significantly enriched among these dolphin PSGs (Table 1), such as GO:0006869: lipid transport, GO:0010876: lipid localization, GO:0007605: sensory perception of sound, GO:0016887: ATPase activity, GO:0006096: glycolysis, GO:0006099: tricarboxylic acid cycle, GO:0050917: sensory perception of umami taste, GO:0003012: muscle system process, and GO:0003774: motor activity. These functional clusters might be related to fat storage, echolocation, energy metabolism, and locomotion in cetaceans.

In order to investigate the tissue-specificity of these PSGs, we utilized the Uniprot tissue (UP_tissue) annotation database. All of the clustered tissues are reported in Figure 3, from which we found that the PSGs are largely expressed in tissues involved with the nervous (199 genes in brain, 20 genes in amygdala and 5 genes in peripheral nervous system), reproductive (99 genes in testis and 82 genes in placenta), and immune systems (28 genes in spleen). In addition to these tissues which commonly express PSGs in many mammalian lineages (Kosiol et al. 2008), we also found dolphin PSGs that are expressed in the kidney (ZC3H11A, FRAS1, FREM1, FREM2, LIMCH1, DNAH7, PEG3, and ARID5B), nasal polyp (DNAH3, DNAH1, and DNAH7) and salivary gland (MYO7B, RRM2B, CDH24, LIMCH1, DSC2, C1orf168, TEP1, SEMA4A, and SIPAIL2) (Figure 3), suggesting potential functional roles for these tissues in processes requiring aquatic adaptation, such as body fluid equilibrium, breath, and digestion and the absorption of food. A detailed list of the functional categories of the PSGs is available in Supplementary Table 2.

We performed an additional multiple test correction for the PSGs according to the “False Discovery Rate” method of Benjamine and Hochberg (1995), even though all of the PSGs have passed a likelihood ratio test (LRT) with raw \(P\) values < 0.01.
Briefly, the \( P \) values of all genes were first ranked from smallest to the largest, and then each \( P \) value was multiplied by the total number of genes (11838) divided by its rank. After correction, 44 and 101 PSGs were retained at 1% and 5% FDR levels, respectively. Functional classification of the retained PSGs showed significant GO term enrichment in categories such as ATPase activity, motor activity, at both FDR levels, and the tissue expression patterns of these genes were similar to that obtained above with the complete gene list (available in Supplementary Table 2).

When the branch-site test for positive selection is significant, then the BEB procedure (Yang, Wong, Nielsen 2005) can be used to calculate the posterior probability of specific sites (codons) being under positive selection. Within the dolphin, PSSs were detected in 125 PSGs. To estimate the effect of sequencing errors, in the low-coverage genome, on our identified PSSs, we randomly chose 48 PSGs that hold one or more PSSs (52 sites in total) and re-sequenced the candidate gene regions. After comparing our newly generated sequences to the reference dolphin genome, we found only 2 sites that showed inconsistent base calling, suggesting that our false-positive rate is only 3.8% (2/52). Based on this result, we expect that the sequence quality of dolphin genome is at a high level, despite the fact that the coverage is only 2.59×. The average size of the 48 amplified gene regions (containing the 52 resequenced sites) was about 390bp (see Supplementary Table 1), thus the sequencing error rate (assuming the all inconsistent sites are errors and not just polymorphisms) of the dolphin genome could be estimated to be 0.1% (2/(48*390)). Given the potentially high quality of the dolphin genome, and the accurate alignments generated in our analyses, this should have improved our identification of PSGs and PSSs, and resulted in a low number of false-positives. A full list of genes with PSSs is available in Supplementary Table 2.

To understand whether the identified PSSs have potential impact on protein structure or function, we searched some of the PSSs against the InterPro database (Apweiler et al. 2001). We first examined the PSGs related to GO:0006869~lipid transport, where only 3 genes (\( APOA2 \), \( ANXA1 \), and \( ATP8B2 \)) showed evidence of PSSs. As shown in Table 2, we found that the PSSs in each of these 3 genes are
located in one or more of the functional domains. We then analyzed another 20 randomly chosen PSGs and found that 15 of them also have PSSs located in their conserved domains (Table 2).

The results obtained in dolphin were compared to those for the cow. Positive selection was detected in the cow lineage, where 242 (2% of the genes) showed significant evidence of positive selection, a number that is smaller than seen in the dolphin. A larger number of genes experiencing positive selection in the dolphin is consistent with our above result that the mean $dN/dS$ in dolphin is much higher than that seen in the cow (Figure 2). For the cow genes, we also performed a functional clustering analysis using the DAVID tool. While there was some overlap in the biological process and molecular function groups significantly enriched, the most enriched cow PSG categories focused on GO:0006811~ion transport, GO:0055085~transmembrane transport, and GO:0042626~ATPase activity, coupled to transmembrane movement of substances (Supplementary Table S2), a set that is extremely different from the GO terms enriched in the dolphin.

We also compared our results with those of a recent similar study performed by McGowen et al. (McGowen, Grossman, Wildman 2012), who utilized the branch model, instead of the branch-site model, to detect PSGs in dolphin. The branch model identifies genes that have $dN/dS > 1$ as being PSGs. When we compared our results to those from McGowen et al. (McGowen, Grossman, Wildman 2012), we found two major differences: (1) Some of the genes identified as having high $dN/dS$ by McGowen did not yield similar values, with the same model, in our analyses. Examples of these genes are: SULT2B1, FAM174B, KCNJ2, and CRYGN. A major reason for this difference likely is that different alignments were used. Many factors influence alignment generation, including homology prediction, transcript choice, aligner choice, and species usage. In our study, all of the species analyzed are mammals, while those used by McGowen et al. ranged more widely and included the chicken. Moreover, McGowen used the program Muscle to align the genes, while, we utilized Prank, a program that is reported to be a much more accurate aligner (Fletcher, Yang 2010). The use of different primary treatments of the genomic data might cause
differences in the estimation of selective constraints acting on genes. (2) Some of the genes identified by McGowen et al. that had high \( dN/dS \) did not pass the likelihood ratio test during our positive selection detection. Examples of these genes include: \textit{SULT2B1, CCL24, BAG2, TSPO2,} and \textit{THAP1}. This observation might be a result of the genes having a high \( dN/dS \) ratio due to the relaxation of purifying selection, rather than being due to positive selection (Cai, Petrov 2010). Although the gene sets obtained by McGowen et al. (McGowen, Grossman, Wildman 2012) are different from those of our study, the Go term enrichment by PSGs showed fewer differences between the two studies (Supplementary Table 2). McGowen et al. (McGowen, Grossman, Wildman 2012) also identified genes related to lipid metabolism, lung development, and ATPase activity as being enriched in their analyses, although with a different gene list from that of our analyses.

Adaptive evolution involves more than just changes to amino acid sequence of proteins. Changes in expression, and gain and loss of genes, are also involved in adaptive evolution. While we could not directly test changes in gene expression levels, we could examine change in expression, we did examine whether change in gene family size had occurred. Change in gene family size could affect both expression level and gain and loss of genes. To address this we performed an analysis of the size of gene families among our five studied species. According to the Ensembl annotation, and classification of each protein-coding gene, we estimated the size of each gene family in each species and identified gene families that showed expansion and/or contraction specific to the dolphin lineage. Only a few expansion events (eg. ENSFM00250000002216, ENSFM00540000720241, and ENSFM00550000743164 families, available in the Supplementary Table S3) were found to have occurred in dolphin lineage, while the contractions in family size mainly occurred to genes in the “Olfactory receptor” gene families (eg. ENSFM0025000000020, ENSFM00320000100072, and ENSFM00430000230074 families).

Discussion

In this study, we have conducted a comparative genomic analysis of the dolphin
genome in an attempt to identify the underlying genetic mechanisms for aquatic adaptation in a mammal. We detected 368 positively selected genes and 1,238 positively selected sites in the dolphin lineage, which showed significantly enrichment in functions related to specific traits in cetaceans such as fat storage, muscle contraction, sensory perception of sound, and ATPase activity.

**Validation of the low-coverage genome**

Although the current release of the dolphin’s genome has only 2.59× coverage, we discovered few sequencing errors in the dolphin genes. Comparing our amplified sequences to the draft genome sequence, we calculated an maximal error rate of approximately 0.9 bases per kilobase, a rate similar to that inferred by McGowen et al. (McGowen, Grossman, Wildman 2012). Therefore, since the dolphin genome appears to be of high quality, sequencing errors should only have a slight influence the detection of positive selection based on the site sensitive method—branch site model. Combined with our strict alignment generation, including the use of Prank and Gblocks, the final alignments should be of high reliability with few alignment errors.

**Adaptive evolution of lipid metabolism genes and the evolution of fat storage in the dolphin**

Most cetaceans have a thick layer of blubber, which acts as their primary storage location for fat. The thickness of the blubber in whales is about 20cm, being 10 times greater than that of other Artiodactyls species (Pond 1978). Blubber provides many benefits to marine mammals, including saving energy by adding buoyancy while they swim. In addition, blubber is also an efficient thermal insulator. People are increasingly interested in genes that are responsible for fat storage (Sohle et al. 2012), since a growing challenge for health care is obesity (Prentice 2006). Here, we have identified some genes that have adaptively evolved and are closely related to lipid metabolism. These adaptively evolving genes belong to the fatty acid/lipid biosynthetic process (e.g. \textit{DGAT1}, \textit{ACSM3}, \textit{ELOVL2}, and \textit{ELOVL5}) and lipid transport and localization (e.g. \textit{APOA2}, \textit{START}, \textit{APOLD1}, and \textit{PLA2G5}).

Fat storage involves the step-wise conversion of exogenous or endogenous fatty acid to diacylglycerol and ultimately triacylglycerols. Fatty acid synthesis is the first
step during fat storage with acetyl-CoA being the key substrate. In this study, we observed an acyl-CoA synthetase gene, ACSM3, had 3 sites (Q158A, S310R, and Y324D) that experienced positive selection in the dolphin. These results suggest that these changes may provide some advantage in generating acetyl-CoA for fat deposit in the dolphin. As shown in Table 2, the 3 PSS sites are located in the AMP-binding domain, a domain responsible for the ligase activity. Moreover, two dolphin fatty acid elongase genes (ELOVL2 and ELOVL5) in dolphin were also identified as having PSSs, with ELOVL2 having one site (M264G) and ELOVL5 having three (T21P, N136Y, and H146F).

Diglyceride acyltransferase (DGAT) is an important protein that catalyzes the formation of triglycerides from diacylglycerol and Acyl-CoA, and whose reaction is the terminal step in triglyceride synthesis. DGAT is essential for the formation of adipose tissue and was also identified as a PSG in the dolphin. In addition, several apolipoprotein genes (APOA2, APOLD1, and LRP1) were also observed to have experienced positive selection, potentially supporting adaptation of lipid transport and localization for fat storage. There is evidence that these genes have roles in elevating lipid content and fat accumulation (Warden et al. 1993; Castellani, Goto, Lusis 2001; Castellani et al. 2008; Terrand et al. 2009).

**Adaptive evolution of locomotion-related genes in cetaceans**

Due to the relatively higher density of water compared to air, it is obvious that cetaceans must confront more resistive drag while swimming than other mammals experience during running. To address this issue, cetaceans must use more energy to support muscle contraction. Here, we have observed that 12 and 8 PSGs belong to genes involved in motor activity and muscle contraction, respectively, identifying candidate genes involved in the adaptation of locomotion physiology (Table 1).

Muscle contraction is triggered by the flow of specific ions, including the transport of sodium ions into muscle cells. SCN4A, a member of the sodium channel family, plays a key role in the ability of a cell to generate and transmit electrical signals. Previous clinical medicine studies have shown that pathogenic mutations in SCN4A lead to myotonic spasms and hyperkalemia periodic paralysis (McClatchey et
al. 1992; Sternberg et al. 2001). Adaptive evolution of the SCN4A gene might help dolphins attain high-speed locomotion. Moreover, some actin cytoskeleton (GO:0015629~actin cytoskeleton) and myosin complex (GO:0016459~myosin complex) proteins, key proteins in the muscle system, were also detected to have undergone positive selection.

Adaptation to locomotion in water was essential for foraging by cetaceans, as they need to dive to great depths and for a long time duration predation. This form of locomotion has large energy costs (Croll et al. 2001), thus, cetaceans must make full advantage of aspired oxygen to generate energy. In this study, genes for the aerobic (TCA cycle, \textit{CS}, \textit{SDHA}, and \textit{MDH1B}) and anaerobic (glycolysis, \textit{HK1}, \textit{OGDH}, \textit{PGAM2}, and \textit{PFKFB1}) respiration were identified to have undergone positive selection in dolphin. The genes \textit{CS} and \textit{HK1} are also key rate limiting enzymes in energy production; therefore, these positively selected genes may have provided some energetic support for the endurance of long dives by cetaceans.

\textbf{Genes for other cetacean-specific traits}

In addition the above PSGs potentially involved in fat storage and locomotion, we also identified genes that were enriched in other functions, such as echolocation and dietary change (Walker, Potter, Macko 1999). As shown in Table 1, 7 PSGs were enriched in GO:0007605~sensory perception of sound category (Table 1), with the genes \textit{CHD7}, \textit{BARHL1}, and \textit{SLC12A2} being closely associated with hearing (Dixon et al. 1999; Li et al. 2002; Vissers et al. 2004). Although the function of these genes in echolocation in the dolphin is unclear, they might provide clues to our understanding of the origin of echolocation in cetaceans.

Unlike their close relatives the Artiodactyls, cetaceans have gradually shifted from an herbivorous to a carnivorous diet, with their major food being fish and crustaceans (Walker, Potter, Macko 1999). The principal nutritional components of these food sources are protein and lipid. In our analysis of the distribution of expression patterns of PSGs we found several genes that are highly expressed in the salivary gland (Figure 3). Combining this observation with the DAVID functional clustering results, where some of these genes enriched are in category
GO:0008236~serine-type peptidase activity, we expect that these adaptively evolving genes may be in part responsible for the shift in diet during cetacean evolution.

An analysis of the gene family evolution was also conducted in this study, as change in gene number is an additional molecular mechanism underlying adaptive evolution (Zhang et al. 2002). While only a few expansion events were occurred in the dolphin lineage (Supplementary Table S3), some have adaptive potential. The dynein domain family (ENSM0000000743164) may be an example, as it is larger in the dolphin than any of the other species examined. Dynein domain containing genes are involved in muscle function, thus adaptive changes in gene number may have supported adaptation to locomotion in water by cetaceans. Moreover, contraction events were also observed in the dolphin lineage, and these were enriched in “Olfactory receptor” gene families (Supplementary Table S3). This observation is consistent with previous finding that some genes involved in pheromonal olfaction in cetaceans had been pseudogenized (Yu et al. 2010). However, these results must be treated with caution, as the current low coverage genome of the dolphin limits the analyses of gene family evolution (Milinkovitch et al. 2010), and thus, an updated high-quality genome would be necessary to conduct further analyses of the genes involve in the adaptation of cetaceans to the aquatic environment.

**Conclusion**

We conducted a genome-wide scan for positively selected genes and sites to investigate the genetic bases of aquatic adaptation by the dolphin. Although a limitation of this approach is the low coverage of the dolphin genome, we utilized strict filters during alignment reconstruction to improve the reliability of the final alignments. These scans, using the improved branch-site model in PAML, revealed 368 PSGs and 1,238 PSSs in the dolphin lineage. The reliability of our results was confirmed by our re-sequencing data, which showed that 50 of 52 randomly chosen PSSs (belonging to 48 PSGs) could be replicated. Functional clustering analysis showed that the PSGs or genes with PSSs were significantly enriched for functions related to fat storage, muscle contraction, ATP generation, perception of sound, and
diet transformation. This study greatly adds to our understanding of the molecular landscape of aquatic adaptation. The low coverage of the dolphin genome, however, limits the detection of other types of genetic mechanism involved in adaption, for example, evolution of gene families and regulatory elements, thus an updated high coverage genome of the dolphin would greatly help to complement the current analyses and allow a better understanding of the aquatic adaptations of cetaceans. Along with the development of the genome sequencing technologies, the sequencing of the genomes of additional aquatic mammals (including Pinnipeds and Sirenia) should provide more resources for investigating convergent or parallel evolution with respect to aquatic adaptation in mammals.

Acknowledges

This research was supported by grants from the National Natural Science Foundation of China (2007CB411600, 2008GA001) and the Bureau of Science and Technology of Yunnan Province (31061160189).

References


**Table 1.** Some functional categories enriched by dolphin positively selected genes.

<table>
<thead>
<tr>
<th>Biological process</th>
<th>Gene number</th>
<th>P value</th>
<th>Fold Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0051693–actin filament capping</td>
<td>4</td>
<td>0.010194345</td>
<td>8.722114765</td>
</tr>
<tr>
<td>GO:0007605–sensory perception of sound</td>
<td>7</td>
<td>0.015587711</td>
<td>3.461870293</td>
</tr>
<tr>
<td>GO:0043588–skin development</td>
<td>4</td>
<td>0.021749614</td>
<td>6.616776718</td>
</tr>
<tr>
<td>GO:003012–muscle system process</td>
<td>9</td>
<td>0.024144751</td>
<td>2.569908815</td>
</tr>
<tr>
<td>GO:0006869–lipid transport</td>
<td>8</td>
<td>0.031724331</td>
<td>2.646710687</td>
</tr>
<tr>
<td>GO:0006936–muscle contraction</td>
<td>8</td>
<td>0.040603444</td>
<td>2.508320586</td>
</tr>
<tr>
<td>GO:0003012–muscle system process</td>
<td>8</td>
<td>0.044559503</td>
<td>2.444143267</td>
</tr>
<tr>
<td>GO:0006096–glycolysis</td>
<td>4</td>
<td>0.073721144</td>
<td>4.082692018</td>
</tr>
<tr>
<td>GO:0006099–tricarboxylic acid cycle</td>
<td>3</td>
<td>0.08174332</td>
<td>6.257169288</td>
</tr>
<tr>
<td>GO:0046356–acetyl-CoA catabolic process</td>
<td>3</td>
<td>0.08174332</td>
<td>6.257169288</td>
</tr>
<tr>
<td>GO:0045927–positive regulation of growth</td>
<td>5</td>
<td>0.084919683</td>
<td>2.998226950</td>
</tr>
<tr>
<td>GO:0010876–lipid localization</td>
<td>8</td>
<td>0.0996468</td>
<td>19.18865248</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Molecular function</th>
<th>Gene number</th>
<th>P value</th>
<th>Fold Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0003774–motor activity</td>
<td>12</td>
<td>1.84E-04</td>
<td>4.048542176</td>
</tr>
<tr>
<td>GO:0003779–actin binding</td>
<td>18</td>
<td>4.79E-04</td>
<td>2.645213139</td>
</tr>
<tr>
<td>GO:0016887–ATPase activity</td>
<td>15</td>
<td>0.010251971</td>
<td>2.151545612</td>
</tr>
<tr>
<td>GO:0005319–lipid transporter activity</td>
<td>5</td>
<td>0.041911263</td>
<td>1.373623674</td>
</tr>
<tr>
<td>GO:008034–lipoprotein binding</td>
<td>4</td>
<td>0.035687</td>
<td>5.475171323</td>
</tr>
<tr>
<td>GO:0070325–lipoprotein receptor binding</td>
<td>5</td>
<td>0.037857</td>
<td>9.581549815</td>
</tr>
<tr>
<td>GO:008236–serine-type peptidase activity</td>
<td>8</td>
<td>0.079084823</td>
<td>2.153157261</td>
</tr>
<tr>
<td>GO:042623–ATPase activity, coupled</td>
<td>11</td>
<td>0.058341087</td>
<td>3.802202308</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cellular component</th>
<th>Gene number</th>
<th>P value</th>
<th>Fold Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0034706–sodium channel complex</td>
<td>4</td>
<td>0.003404288</td>
<td>12.71840796</td>
</tr>
<tr>
<td>GO:0016459–myosin complex</td>
<td>5</td>
<td>0.046770479</td>
<td>3.668771527</td>
</tr>
<tr>
<td>GO:0015629–actin cytoskeleton</td>
<td>11</td>
<td>0.056035936</td>
<td>1.950313488</td>
</tr>
<tr>
<td>GO:042383–sarcolemma</td>
<td>5</td>
<td>0.05129551</td>
<td>3.559255959</td>
</tr>
<tr>
<td>Gene</td>
<td>Domain hits</td>
<td>Start</td>
<td>End</td>
</tr>
<tr>
<td>----------</td>
<td>------------------------------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>APOA2</td>
<td>ApoA-II</td>
<td>24</td>
<td>99</td>
</tr>
<tr>
<td>ANXA1</td>
<td>ANNEXIN</td>
<td>96</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>ANNEXIN</td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td>ATP8B2</td>
<td>ATPase_P-type</td>
<td>811</td>
<td>925</td>
</tr>
<tr>
<td>ACSM3</td>
<td>AMP-binding</td>
<td>96</td>
<td>508</td>
</tr>
<tr>
<td>C2orf77</td>
<td>Trichoplein</td>
<td>148</td>
<td>461</td>
</tr>
<tr>
<td>COL3A1</td>
<td>Collagen</td>
<td>959</td>
<td>1015</td>
</tr>
<tr>
<td>EIF2AK2</td>
<td>PROTEIN_KINASE_DOM</td>
<td>267</td>
<td>538</td>
</tr>
<tr>
<td>EYA4</td>
<td>EYA-cons_domain</td>
<td>375</td>
<td>644</td>
</tr>
<tr>
<td>GLB1L2</td>
<td>GLHYDRLASE35</td>
<td>310</td>
<td>325</td>
</tr>
<tr>
<td></td>
<td>Glyco_hydro_35</td>
<td>53</td>
<td>365</td>
</tr>
<tr>
<td>MFSD12</td>
<td>MFS_2</td>
<td>23</td>
<td>422</td>
</tr>
<tr>
<td>NBR1</td>
<td>PB1 domain</td>
<td>4</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>PB1</td>
<td>5</td>
<td>84</td>
</tr>
<tr>
<td>PARP9</td>
<td>Macro</td>
<td>335</td>
<td>446</td>
</tr>
<tr>
<td></td>
<td>Appr-1^-p processing enzyme</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PGAM2</td>
<td>His_Phos_1</td>
<td>5</td>
<td>191</td>
</tr>
<tr>
<td></td>
<td>pgm_1: phosphoglycerate mutase 1 family</td>
<td>5</td>
<td>215</td>
</tr>
<tr>
<td></td>
<td>Phosphoglycerate mutase family</td>
<td>5</td>
<td>193</td>
</tr>
<tr>
<td>RAVER2</td>
<td>RNA recognition motif</td>
<td>143</td>
<td>216</td>
</tr>
<tr>
<td>SBF1</td>
<td>SSF50729</td>
<td>895</td>
<td>1039</td>
</tr>
<tr>
<td>SCN5A</td>
<td>Ion_trans</td>
<td>1241</td>
<td>1469</td>
</tr>
<tr>
<td>SLC20A2</td>
<td>signal-peptide</td>
<td>1</td>
<td>21</td>
</tr>
<tr>
<td>TEX11</td>
<td>SPO22</td>
<td>161</td>
<td>418</td>
</tr>
</tbody>
</table>
**Figure legends**

**Fig. 1** Phylogenetic tree used in this study.

The accepted phylogeny of the species used for the screen for positively selected genes. The bold line represents the dolphin lineage, the lineage where positive selection was detected.

![Phylogenetic tree](image)

**Fig. 2** Comparisons of rates of evolution at silent and replacement sites in the dolphin and cow lineages.

The molecular traits, $dN$, $dS$, and $dN/dS$, on the lineages leading to the dolphin and the cow are presented. For this comparison, genes with $dS < 0.0005$ were excluded.

**Comparisons of molecular traits between dolphin and cow**

![Graph](image)
**Fig. 3** Tissue-specific expression pattern of dolphin PSGs.

Expression of the human orthologs of the dolphin PSGs was used to examine expression patterns. Numbers of PSGs expressed in human tissues is presented. Expression pattern of the human orthologs of the dolphin PSGs were obtained from the DAVID functional annotation. Most of the PSGs are expressed in the brain, a complex tissue, potentially indicating a relationship to the enlarged size of the brain in cetaceans (McGowen, Grossman, Wildman 2012). In addition, dolphin PSGs are also enriched in other tissues, including kidney, motor cortex, and salivary gland.