

Supplementary Materials

Evolution of p53 pathway-related genes provides insights into anticancer mechanisms of natural longevity in cetaceans

Xing Liu¹, Fei Yang¹, Yi Li¹, Zhen-Peng Yu¹, Xin Huang¹, Lin-Xia Sun¹, Wen-Hua Ren¹, Guang Yang¹, Shi-Xia Xu^{1,*}

¹Jiangsu Key Laboratory for Biodiversity and Biotechnology, College of Life Sciences, Nanjing Normal University, Nanjing, Jiangsu 210023, China

*Corresponding author, E-mail: xushixia@njnu.edu.cn

Supplementary Materials and Methods

Identification of long-lived species

In this study, we used two different standards to define long-lived species: MLS and LQ. LQ is the ratio of observed longevity to expected longevity following Austad and Fisher (Austad and Fischer, 1991), which can indicate how long a species' lifespan compares to that of other species of a similar size. The expected longevity was calculated for each species by fitting a linear regression (Schmidt-Nielsen and Knut, 1984): $Y = aM^b$, where Y is the expected MLS, M is BM, a is a constant representing the proportionality coefficient, and b is the scaling exponent. We first obtained MLS and BM records from the AnAge database (Tacutu et al., 2018) for all nonflying eutherian mammals ($n=823$, **Supplementary Table S2**) to calculate the a and b values. This regression equation also applies to cetaceans. The MLS and BM records of 65 cetacean species were collected from the AnAge & Amniote life history databases (Myhrvold et al., 2015) and published books such as Jefferson et al. (2011) (**Supplementary Table S3**). Species with an LQ or MLS greater than 0.5 SD (standard deviations) from the mean of 65 cetaceans were classified as long-lived. To assess whether ancestral nodes in the tree belong to the long-lived species, the ancestral state of MLS and LQ was reconstructed through the maximum-likelihood method implemented in the APE R-package (Paradis et al., 2004). We used the time-calibrated supertree from McGowen et al. (2020) as the reference tree for the cetacean phylogeny.

Identification of multiple copy genes in cetacean genomes

We used 16 cetacean genomes (three mysticetes and 13 odontocetes) for this study with scaffold N50>1 M downloaded from NCBI and the genome of longest-lived bowhead whale from its corresponding resource (<http://www.bowhead-whale.org/>, **Supplementary Table S4**). The reference gene set used in this study came from the human p53 signaling pathway in KEGG (hsa04115). We utilized the method of Tollis et al. (2020) to test if multiple copies involved in the p53 pathway occurred in cetaceans. We first used BLAT v3.4 software (Kent, 2002) to look for p53 pathway orthologous genes in each cetacean genome (minscore=55, minidentity=60), using all the known human protein sequences as initial references. To collect p53 pathway gene paralogs, the putative homologs in cetaceans with a length at least 2/3 that of the coding sequence were taken as queries in blast searches against the human protein database using BLASTx v2.5 (Boratyn et al., 2012), and cetacean copies were only kept when the top hit had over 65% sequence identity to the human sequence. Additionally, genes with multiple copies identified in cetaceans were scanned in the genomes of 17 non-cetacean species to detect whether the duplicated genes were unique to cetaceans (**Supplementary Table S4**). To assess the influence of genome assembly quality on copy number detection, we analyzed the correlation between the normalized copy numbers (the total number of copies divided by the total number of genes) and genome assembly length and scaffold N50.

Sequence retrieval and alignment

We performed TBLASTN v2.5 searches (Altschul et al., 1990) against cetacean genomes using bottlenose dolphins genes as queries to obtain the coding sequence of high-confidence single-copy orthologs of the p53 pathway, setting the expected cut off

value to 1e-5 (**Supplementary Table S6**). The downloaded sequences that covered at least 75% of the entire coding sequences were retained for further analysis. Multiple alignments of orthologous sequences were first carried out with MACSE v2.0, a tool that prevents the disruption of frameshifts and stop codons (Ranwez et al., 2018). Sequences were then aligned with PRANK v.170427, which outperforms other alignment tools and produces the fewest false positives (Löytynoja, 2014). Finally, potentially unreliable regions of multiple alignments were removed using the Gblocks v0.91b program under parameters “-t=c -b1=5 -b2=6 -b3=8 -b4=10 -b5=h” (Castresana, 2000).

Selection detection

To elucidate the molecular evolution of orthologous genes involved in the p53 pathway, we estimated the rates of nonsynonymous (d_N) and synonymous (d_S) ($\omega=d_N/d_S$) using the maximum likelihood (ML) method in the CODEML program implemented in PAML v4.9 (Yang, 2007). Values of $\omega>1$, $=1$, and <1 indicate positive selection, neutral evolution, and purifying selection, respectively. A well-accepted phylogeny of cetaceans (McGowen et al., 2020) was used as an input tree for analyses of each orthologous gene.

To evaluate whether positive selection was restricted to long-lived cetacean lineages, we used the free-ratio model and branch-site model implemented in the CODEML program. The free-ratio model (M1), which allow independent ω values for each branch, was compared with the null one-ratio model (M0) with the same ω for all branches in a tree. The branch-site model, which assumes that codons are under positive selection along a specific lineage with $\omega_2>1$, was compared with a null model Ma0 with a fixed $\omega_2=1$. The positively selected sites were identified using a Bayes Empirical Bayes (BEB) analysis (Yang et al., 2005) with posterior probabilities ≥ 0.80 . A false discovery rate (FDR, cutoff=0.05) correction for multiple tests was conducted in the branch-site model analysis (Anisimova and Yang, 2007). All nested models were compared using the likelihood ratio test (LRT) to determine which models were statistically different from the null model, P values < 0.05 after FDR correction were considered significant. We further manually checked the putative positively selected sites identified by the branch-site model to minimize the impact of potential false positives. The putative positively selected sites were removed if the aligned positions of the positively selected sites were located nearly or within a poorly conserved section, such as close to insertions or deletions or surrounded by large gaps. The positively selected sites further employed TreeSAAP v3.2 to measure the selective influences on 31 structural and biochemical amino acid properties, the residues that had category Z-scores greater than six were regarded as radical amino acid changes (Woolley et al., 2003).

Additional robustness of statistical significance of aBSREL (adaptive branch-site random effects likelihood, Smith et al., 2015) and BUSTED (branch-site unrestricted statistical test for episodic diversification, Murrell et al., 2015) implemented in Datamonkey were further used to investigate whether episodic positive selection acted on the long-lived lineages. aBSREL identifies branches under positive selection without a priori knowledge about which lineages are of interest by sequential likelihood

ratio tests. BUSTED is capable of detecting episodic positive selection that acts on a subset of branches in the phylogeny in at least one site within the alignment (Murrell et al., 2015). BUSTED, which splits branches into the foreground and background partitions, includes three ω classes ($\omega_1 \leq \omega_2 \leq \omega_3$, unconstrained model) and tests for positive selection against a constrained null model ($\omega=1$, disallowing positive selection) on the foreground branches.

Association analysis between ω and longevity-associated traits

PGLS regression was used to further assess the relationships between gene evolution and lifespan associated traits (i.e., MLS, BM, and LQ). The maximum-likelihood method was used as a quantitative measure of phylogenetic correlation (λ). Values of λ range between 0 and 1, with λ close to 0 indicating traits that are phylogenetically independent and λ of 1 or close to 1 indicating that the genes show a strong phylogenetic signal. All analyses were performed in R v3.4.2 (Orme et al., 2018). The gene evolutionary rate, ω value, for each gene was calculated using the free-ratio model implemented in the CODEML program of PAML v4.9. The root-to-tip ω (average ω ratio from the ancestral cetacean to each terminal species tip), which includes gene evolutionary history, is more suitable for regressions against phenotypic data from extant species. The ultrametric tree of 17 cetacean species from McGowen et al. (2020) was used as an input tree. Lifespan variables and root-to-tip ω were \log_{10} transformed to improve normality for the regression analysis.

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*, **215**(3):403-410.
- Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Molecular Biology and Evolution*, **24**(5):1219-1228.
- Austad SN, Fischer KE. 1991. Mammalian aging, metabolism, and ecology: evidence from the bats and marsupials. *Journal of Gerontology*, **46**(2):B47-B53.
- Boratyn GM, Schäffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL. 2012. Domain enhanced lookup time accelerated BLAST. *Biology Direct*, (1):1-14.
- Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**(4):540-552.
- Jefferson TA, Webber MA, Pitman RL. 2011. Marine mammals of the world: a comprehensive guide to their identification. Oxford: Elsevier.
- Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome Research*, **12**(4):656-664.

Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. *Methods in Molecular Biology*, 155-170.

McGowen MR, Tsagkogeorga G, Álvarez-Carretero S, Dos-Reis M, Struebig M, Deaville R, et al. 2020. Phylogenomic resolution of the cetacean tree of life using target sequence capture. *Systematic Biology*, **69**(3):479-501.

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, et al. 2015. Gene-wide identification of episodic selection. *Molecular Biology and Evolution*, **32**(5):1365-1371.

Myhrvold NP, Baldrige E, Chan B, Sivam D, Freeman DL, Ernest SM. 2015. An amniote life - history database to perform comparative analyses with birds, mammals, and reptiles: Ecological Archives E096 - 269. *Ecology*, **96**(11):3109-3109.

Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S, Isaac N, et al. 2018. Caper: Comparative analyses of phylogenetics and evolution in R. Version: 0.5.2. <https://CRAN.Rproject.org/package=caper>

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, **20**(2):289-290.

Ranwez V, Douzery EJ, Cambon C, Chantret N, Delsuc F. 2018. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Molecular Biology and Evolution*, **35**(10):2582-2584.

Schmidt-Nielsen K, Knut S-N. 1984. Scaling: why is animal size so important?. Cambridge: Cambridge university press.

Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky-Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Molecular Biology and Evolution*, **32**(5):1342-1353.

Tacutu R, Thornton D, Johnson E, Budovsky A, Barardo D, Craig T, et al. 2018. Human ageing genomic resources: new and updated databases. *Nucleic Acids Research*, **46**(D1):D1083-D1090.

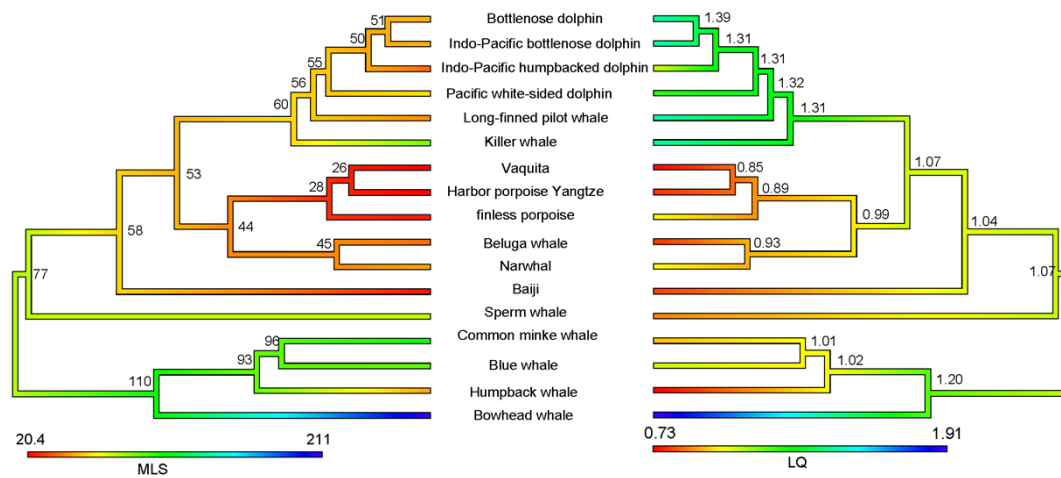
Tollis M, Schneider-Utaka AK, Maley CC. 2020. The evolution of human cancer gene duplications across mammals. *Molecular Biology and Evolution*, **37**(10):2875-2886.

Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA. 2003. TreeSAAP: selection on amino acid properties using phylogenetic trees. *Bioinformatics*, **19**(5):671-672.

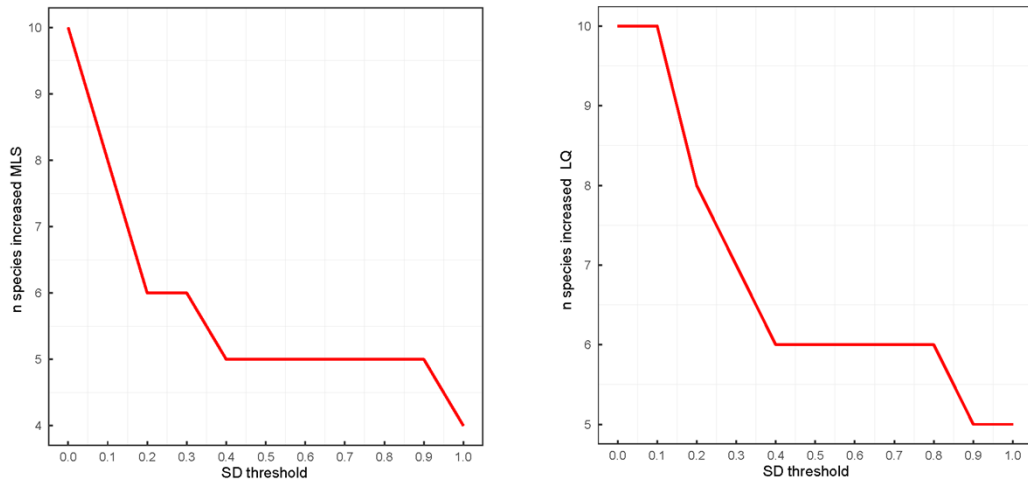
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, **24**(8): 1586-1591.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution*, **22**(4):1107-1118.

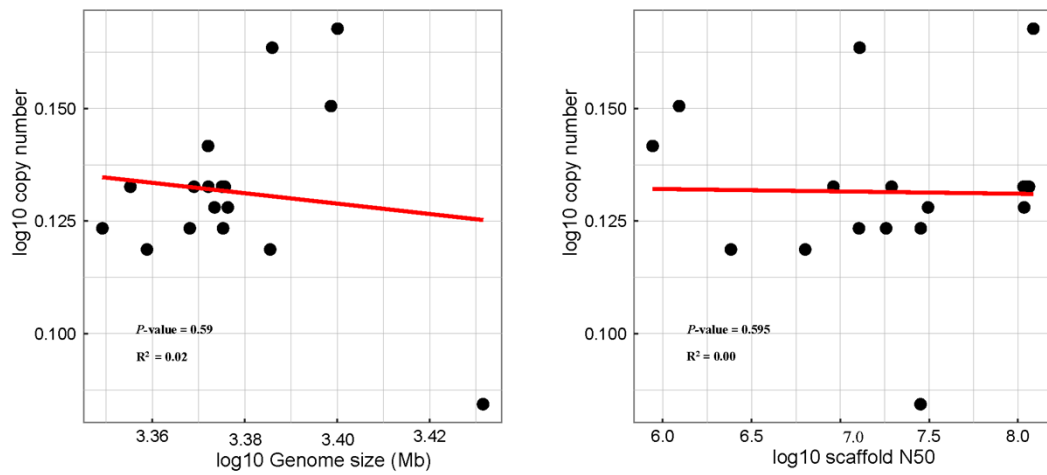
Supplementary Figures



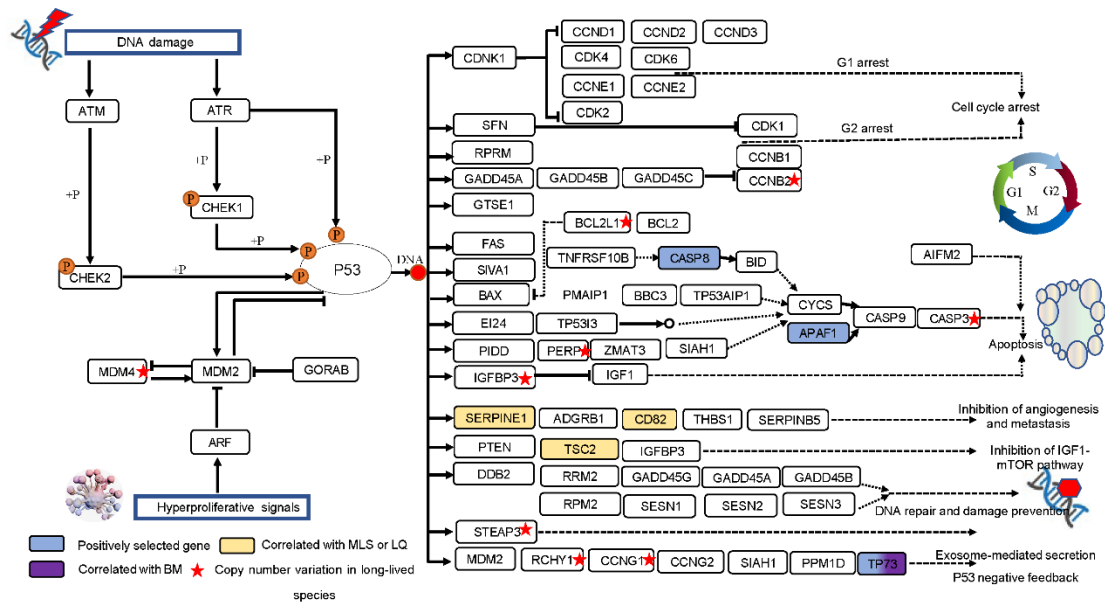
Supplementary Figure S1 Ancestral state reconstruction for LQ and MLS implemented in APE R package



Supplementary Figure S2 Sequential threshold selection ranging from 0 to 1.0 SD from the mean of 17 cetaceans



Supplementary Figure S3 Estimates of copy number variants not related to genome assembly length or scaffold N50



Supplementary Figure S4 Diagram of p53 pathway-related gene evolution in long-lived cetaceans

Supplementary Tables

Supplementary Table S1 List of gens in the p53 signaling pathway

Supplementary Table S2 MLS and BM records of 823 nonflying eutherian mammals from the AnAge database

Supplementary Table S3 Phenotypic data used to calculate mean lifespan of 65 cetaceans

Supplementary Table S4 Genome version and phenotypic data of 17 cetaceans and 17 mammals used in this study

Supplementary Table S5 Copy number variation of p53 pathway genes in cetacean species and mammals

Supplementary Tables S1–S5 are listed as a separate excel file due to their large size

Supplementary Table S6 GenBank accession numbers for the *Tursiops truncatus*

genes used for TBLASTN

Gene	Accession numbers
<i>ADGRB1</i>	XM_033842849.1
<i>AIFM2</i>	XM_033841541.1
<i>APAF1</i>	XM_019918393.2
<i>ATM</i>	XM_033862082.1
<i>ATR</i>	XM_019934297.2
<i>BAX</i>	XM_033845723.1
<i>BBC3</i>	XM_033844600.1
<i>BID</i>	XM_033867288.1
<i>CASP8</i>	XM_019939139.2
<i>CASP9</i>	XM_033842913.1
<i>CCND1</i>	XM_033861345.1
<i>CCND2</i>	XM_033867173.1
<i>CCND3</i>	XM_033864265.1
<i>CCNE1</i>	XM_019941011.2
<i>CCNE2</i>	XM_033843339.1
<i>CCNG2</i>	XM_033857152.1
<i>CD82</i>	XM_033860501.1
<i>CDKN1A</i>	XM_033864359.1
<i>CDKN2A</i>	XM_033858018.1
<i>CHEK2</i>	XM_033837913.1
<i>COP1</i>	XM_019952340.2
<i>DDB2</i>	XM_004322699.2
<i>FAS</i>	XM_019941497.2
<i>GADD45B</i>	XM_033854871.1
<i>GADD45G</i>	XM_004318303.3
<i>GORAB</i>	XM_004317872.3
<i>GTSE1</i>	XM_019933139.2
<i>IGF1</i>	XM_004323127.3
<i>MDM2</i>	XM_033866334.1
<i>PIDD1</i>	XM_033861465.1
<i>PMAIP1</i>	XM_019937161.2
<i>PPM1D</i>	XM_019945482.1
<i>PTEN</i>	XM_019938503.2
<i>RPRM</i>	XM_004319723.3
<i>SERPINB5</i>	XM_019937194.2
<i>SERPINE1</i>	XM_033840247.1
<i>SESNI</i>	XM_019929540.2
<i>SESNI2</i>	XM_033839721.1

<i>SESN3</i>	XM_033862123.1
<i>SHISA5</i>	XM_019946704.2
<i>SIVA1</i>	XM_004322704.3
<i>TP53</i>	XM_019944223.2
<i>TP53I3</i>	XM_033838209.1
<i>TP73</i>	XM_033844567.1
<i>TSC2</i>	XM_033839628.1
<i>ZMAT3</i>	XM_019946249.2

Supplementary Table S7 Evidence of positive selection in cetaceans using free-ratio and branch-site model analysis
in PAML

Gene	Models	-lnL ^a	2 ΔlnL ^b	ω	P-value ^c	Positive selective branches/site	Radical changes in AA properties ^d
Branch model							
<i>APAF1</i>	one ratio	7335.255476		0.2724	P = 0.008		
	free ratio	7363.029643	55.5483	ω variation for each branch		<i>T. aduncus</i> ; LCA of <i>M. novaeangliae</i> <i>G. melas</i> ; <i>P. sinus</i>	
<i>AIFM2</i>	one ratio	3371.39412		0.3128	P = 0.0410		
	free ratio	3347.203975	48.3803	ω variation for each branch			
<i>CASP8</i>	one ratio	3114.260535		0.4487	P = 0.0005	<i>L. obliquidens</i> ; LCA of delphinids	
	free ratio	3081.177543	66.1660	ω variation for each branch			
Branch-site model							
Foreground branch: <i>G. melas</i>							
<i>AIFM2</i>	Null (Ma0)	3029.974424		ω0 = 0.00000 ω1 = 1.00000 ω2 = 1.00000			
	Alternative (Ma)	3024.030126	11.8885	ω0 = 0.00000 ω1 = 1.00000 ω2 = 999.00000	P = 0.0006	459	R _F ; P _C ; α _C ; K ⁰
Foreground branch: LCA of <i>P. phocoena</i>							
<i>TP53I3</i>	Null (Ma0)	2142.912639		ω0 = 0.00000 ω1 = 1.00000 ω2 = 1.00000			
	Alternative (Ma)	2140.398476	5.0283	ω0 = 0.00000 ω1 = 1.00000 ω2 = 99.13473	P = 0.0249	236 254	α _C ; E _t pHi; Mv; Mw; V ⁰
Foreground branch: <i>B. acutorostrata</i>							
<i>CCNE1</i>	Null (Ma0)	1969.5177		ω0 = 0.00000, ω1 = 1.00000, ω2 = 1.0000			
	Alternative	1966.2617	6.5118	ω0 = 0.00000 ω1 = 1.00000	P = 0.0107	339	NA

TP53	(Ma) Null (Ma0)	2155.1763		$\omega 2 = 999.00000$ $\omega 0 = 0.00000$ $\omega 1 = 1.00000$				
	Alternative (Ma)	2152.7685	4.8155	$\omega 2 = 1.00000$ $\omega 0 = 0.00000$ $\omega 1 = 1.00000$ $\omega 2 = 999.00000$	P = 0.0282	18		NA
Foreground branch: LCA of Phocoenidae and Monodontidae								
GTSE1	Null (Ma0)	5554.4296		$\omega 0 = 0.04091$, $\omega 1 = 1.00000$, $\omega 2 = 1.00000$				
	Alternative (Ma)	5551.5030	5.7850	$\omega 0 = 0.00000$ $\omega 1 = 1.00000$ $\omega 2 = 185.90141$	P = 0.0053	522		pk'; R _a
Foreground branch: <i>L. vexillifer</i>								
SIVA1	Null (Ma0)	1059.1290		$\omega 0 = 0.07849$ $\omega 1 = 1.00000$ $\omega 2 = 1.00000$				
	Alternative (Ma)	1056.7690	4.7198	$\omega 0 = 0.08060$ $\omega 1 = 1.00000$ $\omega 2 = 38.82668$	P = 0.0298	21		B _l
Foreground branch: <i>P. catodon</i>								
TP73	Null (Ma0)	3894.0871		$\omega 0 = 0.00000$ $\omega 1 = 1.00000$ $\omega 2 = 1.00000$				
	Alternative (Ma)	3890.4768	7.2207	$\omega 0 = 0.00000$ $\omega 1 = 1.00000$ $\omega 2 = 20.07134$	P = 0.0072	506		pH _i

Physicochemical amino acid properties available in TreeSAAP are as following: α -helical tendencies (P _{α}); Average # surrounding residues (N_s); β -structure tendencies (P _{β}); Bulkiness (B_l); Buriedness (B_r); Chromatographic index (R_F); Coil tendencies (P_c); Composition (c); Compressibility (K⁰); Equil. Const. – ioniza., COOH (pk'); Helical contact energy (C_a); Hydropathy (h); Isoelectric point (pH_i); Long-range n.b. energy (E_l); Mean r.m.s. fluctuat. displace. (F); Molecular volume (M_v); Molecular weight (M_w); Normal. consensus hydrophob. (H_{nc}); Partial specific volume (V⁰); Polar requirement (P_r); Polarity (p); Power to be – C-term., α -helix (α c); Power to be – middle, α -helix (α m); Power to be – N-term., α -helix (α n); Refractive index (μ); Sh.- & med.-range n.b. energy (E_{sm}); Solvent accessible reduct. ratio (R_a); Surrounding hydrophobicity (H_p); Thermodyn. transfer hydrophob. (H_t); Total n.b. energy (E_t); Turn tendencies (P)

Note: a, ln L is the log-likelihood score; b, twice the difference in ln L between the two models compared; c likelihood ratio test p-values were adjusted for multiple testing with FDR method; d, radical changes in amino acid properties under category 6-8 were detected in TreeSAAP. LCA: last common ancestor.

Supplementary Table S8 Evidence of positive selection using the aBSREL

Gene	Log (L)	AIC-c	Branch	P-value
<i>TP73</i>	-3958.21	7994.54	<i>M. novaeangliae</i>	0.0011
<i>GTSE1</i>	-5646.58	11371.25	LCA of Phocoenidae and Monodontidae	0.0198

Note: LCA: last common ancestor; Log (L): Log likelihood of model fit; AIC: Small-sample corrected Akaike information Score

Supplementary Table S9 Evidence of positive selection using the BUSTED

Gene	Model	log L	#par.	Branch set	$\omega 1$	$\omega 2$	$\omega 3$	P-value
<i>TP73</i>	Unconstrained	-3869.6	59	FG	0.00 (59.75%)	0.00 (26.62%)	8.71 (13.63%)	0.000
				BG	0.00 (0.41%)	0.60 (99.57%)	10000 (0.02%)	
	Constrained	-3877.3	58	FG	0.00 (29.18%)	1.00 (42.64%)	1.00 (28.18%)	
				BG	0.60 (99.46%)	1.00 (0.52%)	10000 (0.02%)	

Note: log L, log-likelihood values, # par., the number of parameters