 doi:10.20944/preprints201911.0214.v1

Unbiasing in the Genome Analysis of Iconic Shark Species

Kazuaki Yamaguchi[1] and Shigehiro Kuraku[1,*]

[1]Laboratory for Phyloinformatics, RIKEN Center for Biosystems Dynamics Research. Kobe, Japan

Author contributions: KY performed data analysis, and all the authors wrote the paper. The authors declare no conflict of interest.
[*]Corresponding: shigehiro.kuraku@riken.jp

Abstract: A previous study involving whole genome sequencing of the white shark suggested unique molecular evolution accounting for gigantism and the enhanced longevity of sharks including positive selection of dozens of protein-coding genes potentially involved in genome stability. We performed a reanalysis on some of the genes and identified serious flaws in their results. In this short article, we scrutinize one of the serious problems we identified, report other concerns, and point out a potential bias in analyzing iconic shark species in general.

Keywords: shark; genome; longevity; gigantism; positive selection

Main text

Previously, the paper by Marra et al. (1) suggested unique molecular evolution accounting for gigantism and the enhanced longevity of sharks. Their findings include positive selection of dozens of protein-coding genes potentially involved in genome stability.

They reported positive selection throughout the *Mdm4* gene in the whale shark lineage. Notably, the whale shark *Mdm4* ortholog sequence used in this analysis (XP_020377040.1 in NCBI; presented in Figure 3 by Marra et al. (1)) seems to harbor a wrong open reading frame (ORF) possibly due to problems in genome assembly. The error causes a remarkable dissimilarity to their orthologs, as well as the curated sequence of this whale shark gene supported by transcript sequencing (Figure 1).

Since one of their main findings is undermined by the simple ORF misidentification, we suspect the validity of their findings for the other genes listed in their Table 1. In the white shark protein-coding sequences supplied by Marra et al. (1), we could not identify *Dtl*, *Coq3*, and *Sirt7* orthologs, for which they claim positive selection, while the ORFs probably used for their analysis were identified in the whole genome sequences by referring to the supplied .gff file. Seriously, of these genes, the coding sequences of *Coq3* and *Sirt7* seem to be erroneously predicted, as shown above

 doi:10.20944/preprints201911.0214.v1

for the whale shark *Mdm4*. Their Dataset S1 frequently exhibits inflated values (e.g., 999) for the ratio of nonsynonymous to synonymous substitutions. It is possible that the inflation is caused by ORF misidentification as shown above for *Mdm4* or inclusion of phylogenetically too distant sequences (e.g., paralogs) or species (e.g., teleost fishes that diverged from chondrichthyan species >400 million years ago (2)). The three chondrichthyan species they included in their analysis diverged more than 150 million years ago (2), leaving long branches in between, to be ideally broken by more closely related species with the genome sequences made available earlier (3). The authors justify their use of the branch-site test (4) by referring to existing literature, but the cited literature reports wound healing of the blacktip reef shark, which is irrelevant to the context.

Moreover, a comparison of genome sizes and repeat abundance in their Figure 1 needs to be presented in a uniform format (e.g., number of decimal places) with citation of original information sources (Figure 2). Also, genome sizes should be presented in more precise values instead of '1' or '3'. The genome size comparison in their Figure 1, featuring its increase in the gigantic shark lineages, is misleading: some shark species with relatively small body sizes have comparable or even larger genome sizes (3) (Figure 2). The findings reported by Marra et al. (1), including the core histone gene counts based on a loose definition, need to be reassessed without any bias that genome analysis on only those iconic shark species should readily account for gigantism and high wound healing capacity.

References

1 Marra NJ, et al. (2019) White shark genome reveals ancient elasmobranch adaptations associated with wound healing and the maintenance of genome stability. *Proc Natl Acad Sci USA* 116(10):4446-4455.

2 Irisarri I, et al. (2017) Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol* 1(9):1370-1378.

3 Hara Y, et al. (2018) Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. *Nat Ecol Evol* 2(11):1761-1771.

4 Zhang J, et al. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* 22(12):2472–2479.

5 Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* 30(4):772-780.

6 Allard M, et al. (1998) *Jeanne Calment: from Van Gogh's time to ours, 122 extraordinary years.* New York, WH Freeman and Company.

7 Morton NE (1991) Parameters of the human genome. *Proc Natl Acad Sci USA* 88(17):7474-7476.

8 Fricke H, et al. (2011) The population biology of the living coelacanth studied over 21 years. *Mar Biol* 158(7):1511-1522.

9 Noonan JP, et al. (2004) Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res* 14(12):2397-2405.
10 Gerhard GS, et al. (2002) Life spans and senescent phenotypes in two strains of Zebrafish (*Danio rerio*). *Exp Gerontol* 37(8-9):1055-1068.
11 Postlethwait J, et al. (1998) The zebrafish genome. in *Methods in cell biology* (Vol. 60: pp. 149-163). Academic Press.
12 Hugg D (1996) MAPFISH georeferenced mapping database. Freshwater and estuarine fishes of North America. *Life Science Software. Dennis O. and Steven Hugg*:1278.
13 Hardie DC, Hebert PD (2004) Genome-size evolution in fishes. *Can J Fish Aquat Sci* 61(9):1636-1646.
14 Ojima Y (1990) Cellular DNA contents of fishes determined by flow cytometry. *La Kromosomo II* 57:1871-1888.
15 Francis MP (1997) Spatial and temporal variation in the growth rate of elephantfish (Callorhinchus milii). *New Zeal J Mar Fresh* 31(1):9-23.
16 Francis MP, Ó Maolagáin C (2019) Growth-band counts from elephantfish *Callorhinchus milii* fin spines do not correspond with independently estimated ages. *J Fish Biol* 95(3):743-752.
17 Venkatesh B, et al. (2005). A compact cartilaginous fish model genome. *Curr Biol* 15(3):R82-R83.
18 Chen WK et al. (2007). Age and growth estimates of the whitespotted bamboo shark, *Chiloscyllium plagiosum*, in the northern waters of Taiwan. *Zool Stud* 46(1):92-102.
19 Perry CT et al. (2018) Comparing length-measurement methods and estimating growth parameters of free-swimming whale sharks (*Rhincodon typus*) near the South Ari Atoll, Maldives. *Mar Freshwater Res* 69(10):1487-1495.
20 Michael SW (2005) *Reef sharks and rays of the world.* ProStar Publications.
21 Hamady LL et al. (2014) Vertebral bomb radiocarbon suggests extreme longevity in white sharks. *PLOS One* 9(1):e84006.
22 Schwartz FJ (1986) Comparisons of karyotypes and cellular DNA contents within and between major lines of elasmobranchs. In *Indo-Pacific fish biology: proceedings of the Second International Conference on Indo-Pacific Fishes* (pp. 148-157). Ichthyological Soc. of Japan.
23 Froese R, Palomares MLD (2000) Growth, natural mortality, length–weight relationship, maximum length and length at first maturity of the coelacanth *Latimeria chalumnae*. *Environ Biol Fishes* 58(1):45-52.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2019 doi:10.20944/preprints201911.0214.v1

Figures

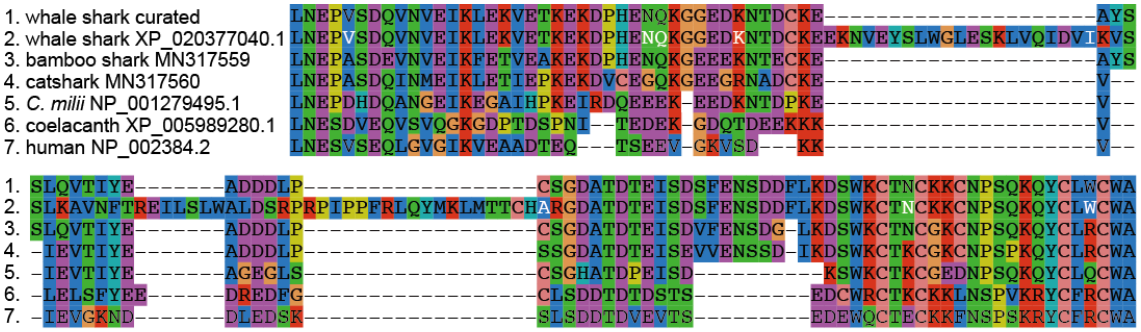
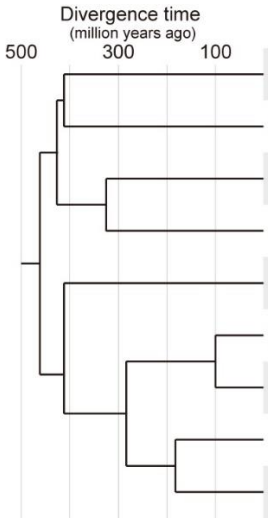


Fig. 1. Multiple alignment of Mdm4 amino acid sequences. This figure shows the amino acid sequence stretch corresponding to the residue 241 to 325 of the human Mdm4 (NP_002384.2 in NCBI) after the alignment using the program MAFFT v7.310 with the option 'linsi' (5). The residues in white letters indicate positively selected residues in the whale shark sequence identified by Marra et al. (1), but most of them are neither unique to the whale shark nor included in the curated whale shark sequence. The GenBank accession IDs for the curated whale shark sequence is MN317558.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 18 November 2019 doi:10.20944/preprints201911.0214.v1



	Lifespan (years)	Genome size (Gbp)	Repeats (%)
human	122.5[6]	3.2[7]*	40.0
coelacanth	48[8]	2.7[9]	43.8
zebrafish	5.5[10]	1.7[11]	56.6
spotted gar	18[12]	1.4[13,14]	19.3
elephant fish (*C.milii*)	20<[15,16]	1.2[17]	41.0
brownbanded bamboo shark	25[18]	4.7[3]	51.8
whale shark	130[19]	3.8[3]	46.6
cloudy catshark	12<[20]	6.7[3]	55.5
white shark	73[21]	6.3[22]	59.2

Fig. 2. Phylogenetic overview of maximum recorded lifespan, genome size, and repeat content in a uniform presentation. This figure shows no obvious trend of the presented genomic factors in the gigantic and long-lived shark lineages. The superscript numbers indicate the literature cited (see References). The branch lengths are proportional to the geological times based on information retrieved from Timescale of Life website (http://www.timetree.org/). For the lifespan of the coelacanth, please also see another reference that proposes the lifespan of >100 years (23). Quantification of the repetitiveness in the genomes was performed as previously described (3). The genome sizes included are based on measurements with flow cytometry except for the human for which a total length of the genome sequences (3.2 Gbp) is conventionally referred to as its genome size (*). Because the lifespan of the brownbanded bamboo shark *Chiloscyllium punctatum* is unavailable, that of a different species in the same genus (*C. plagiosum*) is included.