Commentary: Unbiasing Genome-Based Analyses of Selection: An Example Using Iconic Shark Species

Nicholas J. Marra[1], Michael J. Stanhope[2], Nathaniel K. Jue[3], Vincent P. Richards[4], Stephen J. O'Brien[5,6], Agostinho Antunes[7,8] and Mahmood S. Shivji[5,9]*

[1] Department of Biology, Trustee Science Center, Drury University, Springfield, MO, United States, [2] Population Medicine and Diagnostic Sciences, Cornell College of Veterinary Medicine, Cornell University, Ithaca, NY, United States, [3] School of Natural Sciences, California State University Monterey Bay, Seaside, CA, United States, [4] Department of Biological Sciences, Clemson University, Clemson, SC, United States, [5] Save Our Seas Shark Research Center, Nova Southeastern University, Dania Beach, FL, United States, [6] Guy Harvey Oceanographic Center, Halmos College of Natural Sciences and Oceanography, Nova Southeastern University, Ft. Lauderdale, FL, United States, [7] Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Porto, Portugal, [8] Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal, [9] Guy Harvey Research Institute, Nova Southeastern University, Dania Beach, FL, United States

A Commentary on

Unbiasing Genome-Based Analyses of Selection: An Example Using Iconic Shark Species
by Yamaguchi, K., and Kuraku, S. (2021). Front. Mar. Sci. 8:573853. doi: 10.3389/fmars.2021.573853

Yamaguchi and Kuraku (2021) published an opinion article in this special issue of *Frontiers in Marine Science*, discussing certain analyses of the white shark genome, published earlier in another journal (Marra et al., 2019). Their opinion article involves selected data from our paper, as well as data released by others, subsequent to the analyses reported in our paper. We here, address their opinions.

RESPONSE TO REANALYSIS WITH SEQUENCE CURATION

In our study, we identified 67 genes in the white shark and 21 genes in the whale shark that revealed statistical evidence of positive selection (Marra et al., 2019, **Supplementary Dataset 1**). Yamaguchi and Kuraku (2021) opine that three of these genes, *Mdm4*, *Coq3*, and *Sirt7*, were possibly used without curation and erroneously predicted, which may have resulted in inflated ω values. We note that the species and data in the gene alignments shown by Yamaguchi and Kuraku, to support their view, are not those used in the analysis of Marra et al. Furthermore, the elasmobranch sequence data they include were not available to us at the time we performed the analyses reported in our paper. We are fully aware that erroneous coding sequence prediction is an issue with *ab initio* and computational gene prediction, and in fact so state in our paper (Methods – paragraph 4, section "Positive Selection"). Thus, we conducted best practices consisting of using Gblocks (Talavera and Castresana, 2007) to remove areas of concern from any alignments and manually checking all our alignments that had evidence of selection. Any of these manually checked alignments that had suspect regions or gaps, were edited to correct or remove those regions, and then re-analyzed. Further, we only tested for selection utilizing those portions of the alignments where

there were no gaps or missing data by employing the "clean data" option in PAML (Zhang et al., 2005; Yang, 2007). These curation measures were implemented to minimize as much as possible issues of false positives from erroneous alignments, which, we agree, can falsely inflate evidence of positive selection (Schneider et al., 2009; Markova-Raina and Petrov, 2011). Our approach, including manual curation of the alignments used in our final analyses, was stated in our paper (in the Methods - under the heading "Positive Selection," in the Results – paragraphs 4 and 6 and in the Discussion – paragraph 1).

Yamaguchi and Kuraku focus on *Mdm4*, and it forms a good example to herein discuss some of their opinions in detail. In regards to the analysis of this gene: there is a short stretch of amino acids in the whale shark sequence of *Mdm4* (NCBI accession XP_020377040.1; the NCBI refseq sequence for *Mdm4* for whale shark predicted using the NCBI eukaryotic genome annotation pipeline from an earlier whale shark genome sequence from Read et al., 2017) that was non-homologous to the rest of our sequence alignment. As outlined above, non-homologous (insertion) sequence was excluded in our analysis on the basis of the removal of stretches where missing data occurred in our alignments (detailed in the Methods of Marra et al., 2019). Yamaguchi and Kuraku incorrectly assume (their Figure 1A) that we incorporated the entire erroneous ORF sequence from XP_020377040.1 in our analysis. The authors show an alignment with a different set of taxa than what we used, the above XP_020377040.1 sequence, and their own subsequently curated sequence for this gene from the whale shark (we note that this subsequently curated sequence of Yamaguchi and Kuraku is not listed as the refseq entry from the whale shark genome, and also importantly was only made available on NCBI in November 2019, 9 months after Marra et al. was published). The *Mdm4* sequence alignment (**Supplementary Datafile 1** is the version of the alignment that was analyzed and reported in Marra et al., 2019) that we tested for selection was not only manually curated, but the area with the incorrect "ORF" in their Figure 1A was removed from the whale shark sequence in our alignment prior to the analysis that yielded the result reported in Marra et al. (2019).

Since Yamaguchi and Kuraku created their alignments using not only a different set of species but also unedited NCBI sequences that they assumed were used by Marra et al., they found patterns that would lead to troubling results in a phylogenetic-based analysis of positive selection. We agree that if the analysis was run on the alignment reported by Yamaguchi and Kuraku the results would be erroneous, however this was not what we did, and our analyzed alignment was quite different, thus it is not surprising that they came to a different conclusion.

RESPONSE TO INVESTIGATION OF SUBSTITUTION SATURATION

Marra et al. had anticipated a possible issue of the large divergence time across the species included in the alignments contributing to inflated ω values, and in fact looked for evidence of synonymous saturation (see Marra et al., 2019 Methods – paragraph 4, lines 48–54 of "Positive Selection"). Though we observed some evidence of saturation in the dataset (and identified this in the Methods and Discussion sections of our paper), there are earlier studies indicating that the branch-site test we employed is robust to synonymous saturation and can lead to more false negatives than false positives (Gharib and Robinson-Rechavi, 2013). We included in Marra et al., both the caution that synonymous saturation possibly existed in the dataset and the rationale for our analyses (Methods – paragraph 4, Discussion – paragraph 1). We believe that extreme ω values probably owe more to a small proportion of the alignments contributing to the rate class indicating selection, and as a result having a low or absent synonymous rate (in the case of 999 which is the output value from PAML when there is no denominator in the dn/ds ratio).

CONCLUSIONS

The aim of Marra et al. (2019) was to provide the draft genome of the white shark and begin to glean information about comparative genome evolution in the three chondrichthyan species for which coding sequence data was available at the time of our analysis. One of the opinions of Yamaguchi and Kuraku appears to be that one of our aims was to explain large body size/longevity as a predictor of genome stability, and that in the process we ignored smaller bodied and shorter-lived species. This is not the case; we state in Marra et al. (2019) that the data from the two species in Hara et al. (2018) (the brownbanded bamboo shark and the cloudy catshark), were unavailable *at the time of our analysis* and we acknowledge this circumstance in two places in Marra et al. (2019): in the Introduction – paragraph 2 - "More recently, two additional elasmobranch genomes were published—the brownbanded bamboo shark (*Chiloscyllium punctatum*), and the cloudy catshark (*Scyliorhinus torazame*) (Hara et al., 2018)— these latter two were published only a few weeks before the submission of this report and therefore could not be included in the comparative genomic analyses herein presented"; and in the Methods – paragraph 4 - "The analysis included the genome coding sequences available on GenBank as of September 1, 2017"). In Marra et al. (2019) we took the precautions that we could, in order to take a first look at molecular evolution at a genome level in elasmobranchs, employing the two elasmobranch species with genomes available at the time (white shark and whale shark), both having similar longevity and size and large repeat percentages. As a result, we hypothesized that consistent themes in the data may have a link to those shared life history characteristics. We do not claim in Marra et al. to have done a systematic analysis testing that hypothesis explicitly, rather it was a point of discussion consistent with the data we observed.

In addition to the two shark genomes published by Hara et al. (2018) and a more recent contribution by Zhang et al.

(2020) (white-spotted bamboo shark genome), there are other ongoing elasmobranch genome sequencing projects by several groups (https://vertebrategenomesproject.org). As is typical of the scientific process, future investigations of genome evolution and molecular adaptation in chondrichthyans will certainly benefit from data on genomes from additional divergent species. We look forward to the power that these and other ongoing chondrichthyan genome projects will add to future comparative analyses, serving to increase the knowledge of a fascinating and diverse group of species.

AUTHOR CONTRIBUTIONS

NM, MJS, and MSS wrote the paper with all authors contributing to editing and final preparation of the paper.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmars.2021.696523/full#supplementary-material

Data Sheet 1 | The file contains a .fasta name, but it is in PHYLIP format.

REFERENCES

Gharib, W. H., and Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Mol. Biol. Evol.* 30, 1675–1686. doi: 10.1093/molbev/mst062

Hara, Y., Yamaguchi, K., Onimaru, K., Kadota, M., Koyanagi, M., Keeley, S. D., et al. (2018). Shark genomes provide insights into elasmobranch evolution and the origin of vertebrates. *Nat. Ecol. Evol.* 2, 1675–1686. doi: 10.1038/s41559-018-0673-5

Markova-Raina, P., and Petrov, D. (2011). High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res.* 21, 863–874. doi: 10.1101/gr.115949.110

Marra, N. J., Stanhope, M. J., Jue, N. K., Wang, M., Sun, Q., Bitar, P. P., et al. (2019). White shark genome reveals ancient elasmobranch adaptations associated with wound healing and the maintenance of genome stability. *Proc. Natl. Acad. Sci. U.S.A.* 116, 4446–4455. doi: 10.1073/pnas.1819778116

Read, T. D., Petit, R. A., Joseph, S. J., Alam, M. T., Weil, M. R., Ahmad, M., et al. (2017). Draft sequencing and assembly of the genome of the world's largest fish, the whale shark: *Rhincodon typus* Smith 1828. *BMC Genomics* 18:532. doi: 10.1186/s12864-017-3926-9

Schneider, A., Souvorov, A., Sabath, N., Landan, G., Gonnet, G. H., and Graur, D. (2009). Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol.* 1, 114–118. doi: 10.1093/gbe/evp012

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Bol.* 56, 564–577. doi: 10.1080/10635150701472164

Yamaguchi, K., and Kuraku, S. (2021). Unbiasing genome-based analyses of selection: an example using iconic shark species. *Front. Mar. Sci.* 8:573853. doi: 10.3389/fmars.2021.573853

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi: 10.1093/molbev/msm088

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* 22, 2472–2479. doi: 10.1093/molbev/msi237

Zhang, Y., Gao, H., Li, H., Guo, J., Ouyang, B., Wang, M., et al. (2020). The white-spotted bamboo shark genome reveals chromosome rearrangements and fast-evolving immune genes of cartilaginous fish. *iScience* 23:101754. doi: 10.1016/j.isci.2020.101754