REPLY TO LOCATELLI ET AL.:

Evaluating species-level accuracy of GenBank metazoan sequences will require experts' effort in each group

Matthieu Leray[a], Nancy Knowlton[b,1], Shian-Lei Ho[c], Bryan N. Nguyen[b], and Ryuji J. Machida[c,1]

Biodiversity studies increasingly rely on DNA sequences obtained from the environment (rather than individual organisms) for basic and applied research (1). Species-level assignment of sequences using genetic databases such as GenBank is often desirable (e.g., detecting invasive species, measuring range shifts, or interpreting interaction networks). Thus, Locatelli et al. (2) rightly emphasize the need to evaluate reliability of species annotations for mitochondrial sequences deposited in GenBank and highlight that we did not do so in our recent study (3). While we found relatively few metazoan sequences mislabeled at higher taxonomic levels (<1% even at the genus level), a species-level assessment was not practical given the scale of our analyses (4,714,864 sequences encompassing 15 mitochondrial genes for all metazoans).

Several biological reasons can impede delineation of species boundaries using mitochondrial sequences (2). Some are intrinsic to species and speciation (population genetic structure, incomplete lineage sorting, and mitochondrial DNA introgression via hybridization) and others to the markers themselves (slow mitochondrial molecular evolution in some Porifera, Cnidaria, and Chordata). Thus, biological incongruences between species names and mitochondrial sequences can be challenging to differentiate from technical artifacts such as taxonomic confusion and revision, sample contamination, data entry mistakes, and amplification of pseudogenes.

The analysis of Locatelli et al. (2) was based on 43 well-studied commercial fish species, which should minimize the difficulty of identifying species from mitochondrial sequences. They looked for multiple peaks in distributions of sequence similarity within individual species, which are not expected if species are well-defined and consistently identified. They surprisingly found nonunimodal distributions in 26 of the 43 species, three of which they discuss in detail.

One of these is *Brevoortia tyrannus*, the Atlantic menhaden, for which hybridization has been documented, including the existence of two *B. tyrannus* mitochondrial haplotype clades (4). A second is *Sebastes miniatus*, the vermillion rockfish, for which Hyde and Vetter (5) reported the presence of genetically distinct clades, suggesting the presence of cryptic species. Finally, we reexamined the sequences for *Cupea pallasii*, the Pacific herring; 42 of these appear to be correct while 3 are outliers. One (JQ354055.1) is clearly not *C. pallasii* and was reported to GenBank as likely an error (3) (although GenBank has not yet flagged problematic sequences uncovered during our analyses). The other two (EU200471.1 and EU200487.1) are likely pseudogenes, as mentioned in the GenBank flat file. All three outlier sequences were removed from the curated MIDORI reference dataset built from GenBank (6).

In sum, two of the examples in ref. 2 would be predicted from the literature, and the third should soon be removed from GenBank or detected by standard quality-control steps. We agree with Locatelli et al. (2) on the importance of refining our knowledge, but even their examples confirm that GenBank is a surprisingly reliable resource. Of course, for the much-less-studied invertebrates (i.e., the vast majority of GenBank sequences), identifying errors in species-level annotations will require a tremendous scientific effort involving close collaborations with experts in each group.

[a]Smithsonian Tropical Research Institute, Smithsonian Institution, Panama City 0843-03092, Republic of Panama; [b]National Museum of Natural History, Smithsonian Institution, Washington, DC, 20560; and [c]Biodiversity Research Centre, Academia Sinica, Taipei 115-29, Taiwan

Author contributions: M.L., N.K., B.N.N., and R.J.M. designed research; M.L., B.N.N., and R.J.M. performed research; and M.L., N.K., S.-L.H., B.N.N., and R.J.M. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: knowlton@si.edu or ryujimachida@gmail.com.

First published November 24, 2020.

1 K. Deiner *et al.*, Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Mol. Ecol.* **26**, 5872–5895 (2017).
2 N. S. Locatelli, P. B. McIntyre, N. O. Therkildsen, D. S. Baetscher, GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 32211–32212 (2020).
3 M. Leray, N. Knowlton, S.-L. Ho, B. N. Nguyen, R. J. Machida, GenBank is a reliable resource for 21st century biodiversity research. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22651–22656 (2019).
4 J. D. Anderson, W. J. Karel, Genetic evidence for asymmetric hybridization between menhadens (*Brevoortia* spp.) from peninsular Florida. *J. Fish Biol.* **71**, 235–249 (2007).
5 J. R. Hyde, R. D. Vetter, The origin, evolution, and diversification of rockfishes of the genus *Sebastes* (Cuvier). *Mol. Phylogenet. Evol.* **44**, 790–811 (2007).
6 R. J. Machida, M. Leray, S. L. Ho, N. Knowlton, Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Sci. Data* **4**, 170027 (2017).