frontiers in Genetics

ORIGINAL RESEARCH
published: 06 February 2018
doi: 10.3389/fgene.2018.00030

Detection of Potential Problematic *Cytb* Gene Sequences of Fishes in GenBank

Xiaobing Li[1], *Xuejuan Shen*[1], *Xiao Chen*[2], *Dan Xiang*[3], *Robert W. Murphy*[4] *and Yongyi Shen*[1,3,5]*

[1] College of Veterinary Medicine, South China Agricultural University, Guangzhou, China, [2] College of Marine Sciences, South China Agricultural University, Guangzhou, China, [3] Joint Influenza Research Centre (SUMC/HKU), Shantou University Medical College, Shantou, China, [4] Centre for Biodiversity and Conservation Biology, Royal Ontario Museum, Toronto, ON, Canada, [5] Key Laboratory of Zoonosis Prevention and Control of Guangdong Province, Guangzhou, China

Fishes are, by far, the most diverse group of vertebrates. Their classification relies heavily on morphology. In practice, the correct morphological identification of species often depends on personal experience because many species vary in their body shape, color and other external characters. Thus, the identification of a species may be prone to errors. Due to the rapid development of molecular biology, the number of sequences of fishes deposited in GenBank has grown explosively. These published data likely contain errors owing to invalid or incorrectly identified species. The erroneous data can lead to downstream problems. Thus, it is critical that such errors get identified and corrected. A strategy based on DNA barcoding can detect potentially erroneous data, especially when intraspecific K2P variation exceeds interspecific K2P divergence. Analyses of the most used DNA marker for fishes (mitochondrial *Cytb*) discovers that intraspecific differences of fishes are generally less than 1%, while interspecific differences are generally higher than 10%. Based on this ruler, our analyses identify 1,303 potential problematic *Cytb* sequences of fishes in GenBank and point to taxonomic problems, errors in identification, genetic introgression and other concerns. Care must be taken to avoid the perpetuation of errors when using these available data.

Keywords: sequence error, DNA barcoding, fish, GenBank, *Cytb*

OPEN ACCESS

Edited by:
Naiara Rodriguez-Ezpeleta, Centro Tecnológico Experto en Innovación Marina y Alimentaria (AZTI), Spain

Reviewed by:
Peter John Unmack, University of Canberra, Australia
Andrey Tatarenkov, University of California, Irvine, United States

***Correspondence:**
Yongyi Shen
shenyy@scau.edu.cn

Specialty section:
This article was submitted to Evolutionary and Population Genetics, a section of the journal Frontiers in Genetics

Received: *15 November 2017*
Accepted: *22 January 2018*
Published: *06 February 2018*

Citation:
Li X, Shen X, Chen X, Xiang D, Murphy RW and Shen Y (2018) Detection of Potential Problematic Cytb Gene Sequences of Fishes in GenBank. Front. Genet. 9:30. doi: 10.3389/fgene.2018.00030

INTRODUCTION

The identification of fishes generally relies on morphology and distribution. However, in practice, problems exist due to the great diversity of fishes, small body sizes of many species, poor preservation of individual specimens and other issues. Further, accuracy in the morphological identification of species depends on personal experience. For many species, abiotic factors such as environmental pertubations can affect body shape, skin color and other external characters (Wilkens and Strecker, 2003). These factors inevitably lead to controversy and misidentification.

DNA barcoding uses a short gene segment to identify species (Hebert et al., 2003a, 2004). Generally, mitochondrial *COI* gene is the marker of choice because differences in sequences between species have been well characterized (Hebert et al., 2003b). This method has been applied to the classification of fishes to facilitate the rapid and accurate identification of species and the discovery of the cryptic species (Fields et al., 2015; Bhattacharya et al., 2016). In DNA barcoding,

a short standardized sequence can distinguish individuals of a species because genetic variation between species usually exceeds that within species (Hebert et al., 2003a; Hajibabaei et al., 2007). In such cases, any gene segment can serve to identify species. Potential errors and taxonomic conundrums can be identified when interspecific genetic variation does not exceed that within species.

Because of advances in sequencing technologies, the number of DNA sequences of fishes has increased explosively in GenBank. For example, fishes now have more than 60,000 sequences of mitochondrial cytochrome *b* (*Cytb*) alone in the database, and this representation is ever increasing. Many sequences have been submitted by labs void of taxonomic expertise. Further, sampling error, contamination, hybridization, introgression, and nuclear pseudogenes can also lead to problems and errors. Consequently, any large database likely contains errors and the perpetuation of erroneous data can lead to downstream problems. Thus, it is critical to identify and correct such errors.

The large gap between *Cytb* intra- and interspecies differences is stable. Consequently, the gene has been used widely in systematics and molecular ecology including the identifications of species of chickens, praomyin rodents and gadid fishes, among many others (Kartavtsev, 2011; Nicolas et al., 2012; Yacoub et al., 2015; Fernandes et al., 2017). Many studies on fishes have used *Cytb* sequences for molecular phylogenetics and population analyses. Therefore, we use *Cytb* to test if DNA barcoding can identify potential erroneous sequences of fishes. This approach has the potential to be used universally to improve the quality of publically available data.

MATERIALS AND METHODS

To obtain the maximum number of sequences, we downloaded all 65,326 *Cytb* records for fishes from NCBI. These sequences, which were uploaded by many labs, many of them were incomplete *Cytb* genes, had different lengths and covered different parts of the gene. Therefore, we employed the following trimming steps to standardize these sequences before calculating sequence divergences: (1) flanking regions of *Cytb* were deleted; (2) sequences were aligned using MAFFT (Katoh and Toh, 2010); (3) to obtain the maximum number of homologous sequences, we balanced the maximum length alignment vs. taxonomic coverage to attain the final trimmed dataset for downstream analyses. The trimmed dataset consisted of 35,130 fragments of 918 bp. When we set the complete *Cytb* for *Carassius auratus* GU135519.1 as the standard, the available fragments ranged from 75 to 998 bp.

DAMBE (Xia and Xie, 2001) was employed to detect for nucleotide substitution saturation. Iss < ss.c was statistically significant ($P = 0$), indicating that the nucleotide substitution was not saturated (Xia et al., 2003). Pairwise divergences (Kimura 2-parameter, K2P) of these sequences were calculated using MEGA 6 (Tamura et al., 2013). Then, intraspecific distances greater than 1% and interspecific distances less than 10% were identified as being potentially problematic. Neighbor-joining trees with 1,000 bootstrap replications were constructed using MEGA 6 (Tamura et al., 2013) to visualize similarity and sequence divergence. Sequences with intraspecific K2P divergences greater than interspecific differences were retained for further evaluation.

RESULTS AND DISCUSSION

The compiled a dataset of *Cytb* sequences of fishes from GenBank exhibited a great diversity of lengths. A clear tradeoff existed between maximizing the length of the alignments and taxonomic coverage (Shen et al., 2013). Usable fragment lengths ranged from 55 to 972 bp. Our final dataset consisted of 35,130 fragments of 918 bp. We regarded GenBank accession number GU135519.1 for *Cytb* to be the standard for all comparisons.

The index of substitution saturation (Iss) is significantly less than the critical Iss.c ($P = 0$) (**Table 1**). This result suggests that the nucleotide substitutions are not saturated. The distribution of genetic distances was shown to vary greatly (Johns and Avise, 1998). Notwithstanding, our intraspecific differences generally fall below 1%, while interspecific differences usually exceed 10% (**Figure 1**). The gap suggests that *Cytb* can efficiently distinguish different species of fishes. Some notable exceptions exist. For example, sequences with shallow interspecific divergence (<10%), deep intraspecific divergence (>1%), and interspecific differences that are much less than intraspecific differences constitute potential errors. Based on this ruler, we identify 1,303 potential problematic *Cytb* gene sequences (Table S1).

Shallow interspecific divergence may owe to several possibilities. (1) Species of recent origin should have very shallow interspecific divergence. For example, the K2P divergence between *Comephorusdy bowskii* and *C. baicalensis* is only 0.4–1.0%, and between *Etheostoma kanawhae* and *E. osburni* a mere 0.4–0.7%. These species appear to have recent origins (Syu et al., 1994; Sun et al., 2007; Geiger et al., 2016). (2) MtDNA introgression can lead to shallow interspecific differences. For example, *Melanotaenia misoolensis* (KC133624.1) is very similar to *M. flavipinnis* (0.2–0.3%), and *M. boesemani* (KC133618.1) shows shallow interspecific divergence with *M. ajamaruensis* (0.3–0.4%). Gene introgression via hybridization occurs in rainbowfishes (Unmack et al., 2013). The low genetic divergence between *Chasmistes brevirostris* and *Deltistes luxatus*is (0.8–1%) is also due to introgressive hybridization (Dowling et al., 2016). Introgressive hybridizations were also found in suckers, darters, barbs and so on (Near et al., 2011; Unmack et al., 2014; Bernal et al., 2017; Schmidt et al., 2017). This reason leads to the unexpected shallow interspecific divergence in many fishes. Nuclear sequences would be helpful to classify recent origin or mtDNA introgression. (3) Errors in species identification

TABLE 1 | Test of substitution saturation of *Cytb* sequences of fishes.

	Iss	I ss.cSym	T	DF	P	Iss.cSym	T	DF	P
4	0.298	0.817	29.382	917	0.000	0.785	27.583	917	0.000
8	0.296	0.784	24.705	917	0.000	0.677	19.302	917	0.000
16	0.294	0.766	22.614	917	0.000	0.565	12.988	917	0.000
32	0.299	0.742	20.696	917	0.000	0.431	6.190	917	0.000

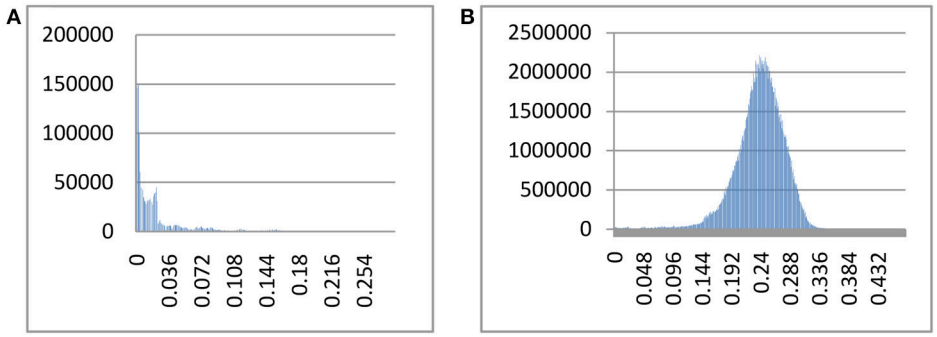

FIGURE 1 | Intra- and interspecific pairwise divergence (Kimura 2-parameter) of *Cytb* in fishes. **(A)** Intraspecific divergence. **(B)** Interspecific divergence.

and conspecificity of the species can also lead to low values of divergence. For example, *Etheostoma spectabile* (FJ381067.1, FJ381066.1, FJ381061.1, and FJ381057.1), *E. bison* (KF377137.1), *E. burri* (FJ381080.1 and AY374262.1), and *E. lawrencei* (KF377157.1 and KF377156.1) show shallow interspecific divergence with *E. caeruleum* (0.3–1.1%). This result suggests conspecificity of the species, or species misidentifications. K2P distances between *Etheostoma sitikuense*, *E. percnurum, E. marmorpinnum* range from 0.2 to 1.1%. The low levels of interspecific divergence indicate either recent divergence or perhaps a taxon-specific slowing of the molecular clock. Although no specific level of divergence can identify species, low interspecific divergence point to a need for further investigation.

Larger than expected intraspecific differences also exist. For example, two sequences of *Paramisgurnus dabryanu* (KM186183.1, KF771003.1) differ from conspecifics by 18.1–19.6%, one *Paracobitis malapterura* (LC167412.1) differs by 22.0–22.4%, two *Etheostoma coosae* (HQ128114.1, AY374266.1) by 10.9–12.2%, two *Rhodeus ocellatus* (KT004415.1, AF051876.1) by 20.0–20.6%, and two *Schizothorax waltoni* (KT833090.1, KT833089.1) by 19.2–20.7%. These cases indicate that at least half of the sequences were either incorrectly identified to species, contamination of DNA occurred in the laboratory, or an errneous sequence was submitted to GenBank.

Species having wide ranges of intraspecific differences are most likely composites of multiple cryptic species. For example, *Etheostoma nigripinne* has complex relationships, and its intraspecific divergences range from 0.0 to 14.5%. Similarly, intraspecific divergences of *E. rufilineatum* range from 0.1 to 12.6%. Many currently recognized species contain a few cryptic species (Köhler et al., 2005; Palandacic et al., 2017; Phuong et al., 2017). Further taxonomic study is necessary for those species with wide ranges of intraspecific differences.

Cases where interspecific differences are much less than intraspecific differences likely owe to problems such as species misidentifications, database errors when submitting sequences to GenBank, laboratory mix-ups, laboratory contamination, and other issues. For example, one sequence of *Etheostoma oophylaxe* (JX547432.1) has shallow interspecific divergence with *E. nigripinne* (0.1–4.1%), but deep intraspecific divergence (13.8–14.5%) (**Figure 2A**). One sequence of *E. artesiae* (HQ128075.1) has relatively low interspecific divergences with *E. swaini* (5.4–7.6%), but deep intraspecific divergence (10.3–10.4%). Three sequences of *E. crossopterum* (JX547246.1; JX547256.1; and JX547253.1) have shallow interspecific divergence with *E. nigripinne* (0–0.3%) but exhibit deep intraspecific divergence (15.6–16.8%). Four sequences of *Acheilognathus signifier* (KF410810.1; KF410811.1; EF483930.1 and JQ714034.1) have low interspecific divergences with *Tanakia koreensis* (0.2–5.2%), yet deep intraspecific divergence (15.4–16.1%; **Figure 2B**). Further investigation into the discordance is desirable.

Other reasons can lead to unexpected values of genetic divergence. (1) Great geographic distances can result in genetic divergence, especially in widely distributed species. (2) Recent origins of species can result in high levels of genetic similarity. (3) Taxonomic change can result in errors. For example, the names *Rutilus lemmingii* and *Chondrostoma lemmingii* differ, but they are the same species, as do *Epinephelus lanceolatus* and *Promicrops anceolatus*. Therefore, we suggest that GenBank (NCBI) provide a mechanism for updating changes in taxonomic classification. (4) Morphologically different species may have essentially identical genes. For example, many species of darters (*Etheostoma*) differ morphologically, but genetically differ slightly. Similarly, *Glossolepis incisus*, *G. pseudoincisus,* and *G. dorityi* are all essentially identical genetically (Unmack et al., 2013). It has to be mentioned that without standard sequences for each species, when two sequences have atypical genetic divergence values, we cannot classify which sequence is correct and which is wrong. Further investigations into species with atypical genetic divergence values (Table S1) can improve the accuracy of the fish mitochondrial database and foster interesting study.

DNA barcoding can complement morphological classifications and provide an alternative approach to assessing species diversity. Now, the approach is widely used to identify species of fishes (Ward et al., 2005; Smith et al., 2008; Ardura et al., 2010; Filonzi et al., 2010). Classifications form the basis of evolutionary research and incorrect taxonomies can negatively affect all other biological investigations. Fishes comprise nearly half of all vertebrate species, and, thus, an accurate classification is essential. Species identification errors in GenBank can mislead

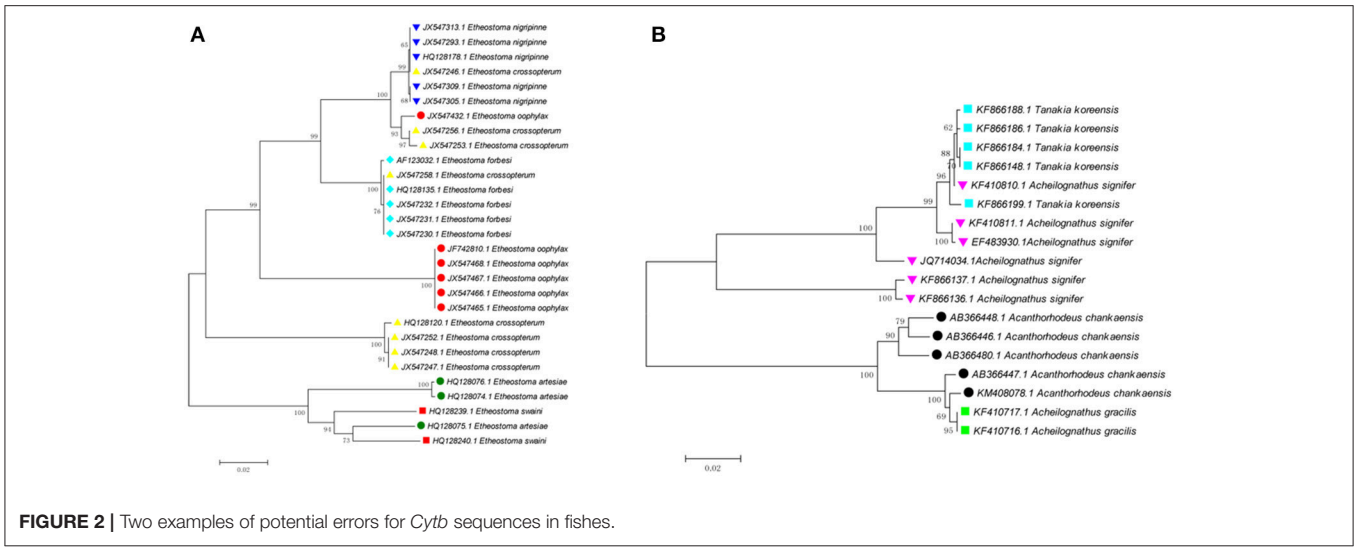

FIGURE 2 | Two examples of potential errors for *Cytb* sequences in fishes.

subsequent research. We detect potentially problematic data for one gene only, *Cytb*, for sequences from fishes. The approach will be useful for other mitochondrial genes and other taxa. DNA barcoding can identify species of fishes, species complexes, sister-species, and discover potentially problematic errors.

AUTHOR CONTRIBUTIONS

XL carried out the data analysis and drafted the manuscript; XS, XC, and DX carried out data analysis; YS designed and coordinated the study, and helped draft the manuscript; RM revised the manuscript. All authors gave final approval for publication.

FUNDING

This work was supported by Guangdong Natural Science Funds for Distinguished Young Scholar (2014A030306046), National Natural Science Foundation of China (41666008), start-up funding from South China Agricultural University for Yongyi Shen, and by a Visiting Professorship for Senior International Scientists from the Chinese Academy of Sciences to RM.

REFERENCES

Ardura, A., Linde, A. R., Moreira, J. C., and Garcia-Vazquez, E. (2010). DNA barcoding for conservation and management of Amazonian commercial fish. *Biol. Conserv.* 143, 1438–1443. doi: 10.1016/j.biocon.2010.03.019

Bernal, M. A., Gaither, M. R., Simison, W. B., and Rocha, L. A. (2017). Introgression and selection shaped the evolutionary history of sympatric sister-species of coral reef fishes (genus: Haemulon). *Mol. Ecol.* 26, 639–652. doi: 10.1111/mec.13937

Bhattacharya, M., Sharma, A. R., Patra, B. C., Sharma, G., Seo, E. M., Nam, J. S., et al. (2016). DNA barcoding to fishes: current status and future directions. *Mitochondrial DNA* 27, 2744–2752. doi: 10.3109/19401736.2015.1046175

Dowling, T. E., Markle, D. F., Tranah, G. J., Carson, E. W., Wagman, D. W., and May, B. P. (2016). Introgressive hybridization and the evolution of lake-adapted Catostomid fishes. *PLoS ONE* 11:e0149884. doi: 10.1371/journal.pone.0149884

Fernandes, T. J., Costa, J., Oliveira, M. B., and Mafra, I. (2017). DNA barcoding coupled to HRM analysis as a new and simple tool for the authentication of Gadidae fish species. *Food Chem.* 230, 49–57. doi: 10.1016/j.foodchem.2017.03.015

Fields, A. T., Abercrombie, D. L., Eng, R., Feldheim, K., and Chapman, D. D. (2015). A novel mini-DNA barcoding assay to identify processed fins from internationally protected shark species. *PLoS ONE* 10:e0114844. doi: 10.1371/journal.pone.0114844

Filonzi, L., Chiesa, S., Vaghi, M., and Nonnis Marzano, F. (2010). Molecular barcoding reveals mislabelling of commercial fish products in Italy. *Food Res. Int.* 43, 1383–1388. doi: 10.1016/j.foodres.2010.04.016

Geiger, M. F., Schreiner, C., Delmastro, G. B., and Herder, F. (2016). Combining geometric morphometrics with molecular genetics to investigate a putative hybrid complex: a case study with barbels *Barbus* spp. (Teleostei: Cyprinidae). *J. Fish. Biol.* 88, 1038–1055. doi: 10.1111/jfb.12871

Hajibabaei, M., Singer, G. A. C., Hebert, P. D. N., and Hickey, D. A. (2007). DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends Genet.* 23, 167–172. doi: 10.1016/j.tig.2007.02.001

Hebert, P. D., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003a). Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270, 313–321. doi: 10.1098/rspb.2002.2218

Hebert, P. D., Ratnasingham, S., and deWaard, J. R. (2003b). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proc Biol Sci.* 7(Suppl. 1), S96–S99. doi: 10.1098/rsbl.2003.0025

Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004). Identification of Birds through DNA Barcodes. *PLoS Biol.* 2:e312. doi: 10.1371/journal.pbio.0020312

Johns, G. C., and Avise, J. C. (1998). A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene. *Mol. Biol. Evol.* 15, 1481–1490. doi: 10.1093/oxfordjournals.molbev.a025875

Kartavtsev, Y. P. (2011). Divergence at Cyt-b and Co-1 mtDNA genes on different taxonomic levels and genetics of speciation in animals. *Mitochondrial DNA* 22, 55–65. doi: 10.3109/19401736.2011.588215

Katoh, K., and Toh, H. (2010). Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* 26, 1899–1900. doi: 10.1093/bioinformatics/btq224

Köhler, J., Vieites, D. R., Bonett, R. M., García, F. H., Glaw, F., Steinke, D., et al. (2005). New amphibians and global conservation: a boost in species discoveries in a highly endangered vertebrate group. *Bioscience* 55, 693–696. doi: 10.1641/0006-3568(2005)055[0693:NAAGCA]2.0.CO;2

Near, T. J., Bossu, C. M., Bradburd, G. S., Carlson, R. L., Harrington, R. C., Hollingsworth, P. R. Jr., et al. (2011). Phylogeny and temporal diversification of darters (Percidae: Etheostomatinae). *Syst. Biol.* 60, 565–595. doi: 10.1093/sysbio/syr052

Nicolas, V., Schaeffer, B., Missoup, A. D., Kennis, J., Colyn, M., Denys, C., et al. (2012). Assessment of three mitochondrial genes (*16S, Cytb, CO1*) for identifying species in the Praomyini tribe (Rodentia: Muridae). *PLoS ONE* 7:e36586. doi: 10.1371/journal.pone.0036586

Palandacic, A., Naseka, A., Ramler, D., and Ahnelt, H. (2017). Contrasting morphology with molecular data: an approach to revision of species complexes based on the example of European Phoxinus (Cyprinidae). *BMC Evol. Biol.* 17:184. doi: 10.1186/s12862-017-1032-x

Phuong, M. A., Bi, K., and Moritz, C. (2017). Range instability leads to cytonuclear discordance in a morphologically cryptic ground squirrel species complex. *Mol. Ecol.* 22:14238. doi: 10.1111/mec.14238

Schmidt, R. C., Bart, H. L. Jr., and Nyingi, W. D. (2017). Multi-locus phylogeny reveals instances of mitochondrial introgression and unrecognized diversity in Kenyan barbs (Cyprininae: Smiliogastrini). *Mol. Phylogenet. Evol.* 111, 35–43. doi: 10.1016/j.ympev.2017.03.015

Shen, Y. Y., Chen, X., and Murphy, R. W. (2013). Assessing DNA barcoding as a tool for species identification and data quality control. *PLoS ONE* 8:e57125. doi: 10.1371/journal.pone.0057125

Smith, P. J., McVeagh, S. M., and Steinke, D. (2008). DNA barcoding for the identification of smoked fish products. *J. Fish Biol.* 72, 464–471. doi: 10.1111/j.1095-8649.2007.01745.x

Sun, Y. H., Xie, C. X., Wang, W. M., Liu, S. Y., Treer, T., and Chang, M. M. (2007). The genetic variation and biogeography of catostomid fishes based on mitochondrial and nucleic DNA sequences. *J. Fish Biol.* 70, 291–309. doi: 10.1111/j.1095-8649.2007.01453.x

Syu, S., Pavlova, M. E., and Belikov, S. I. (1994). Analysis of tandem DNA repeats of cottoid fish in Lake Baikal by direct consensus sequencing. *Mol. Marine Biol. Biotechnol.* 3, 301.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 30, 2725. doi: 10.1093/molbev/mst197

Unmack, P. J., Allen, G. R., and Johnson, J. B. (2013). Phylogeny and biogeography of rainbowfishes (Melanotaeniidae) from Australia and New Guinea. *Mol. Phylogenet. Evol.* 67, 15–27. doi: 10.1016/j.ympev.2012.12.019

Unmack, P. J., Dowling, T. E., Laitinen, N. J., Secor, C. L., Mayden, R. L., Shiozawa, D. K., et al. (2014). Influence of introgression and geological processes on phylogenetic relationships of Western North American mountain suckers (Pantosteus, Catostomidae). *PLoS ONE* 9:e90061. doi: 10.1371/journal.pone.0090061

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., and Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 360, 1847–1857. doi: 10.1098/rstb.2005.1716

Wilkens, H., and Strecker, U. (2003). Convergent evolution of the cavefish Astyanax (Characidae, Teleostei): genetic evidence from reduced eye-size and pigmentation. *Biol. J. Linn. Soc. Lond.* 80, 545–554. doi: 10.1111/j.1095-8312.2003.00230.x

Xia, X., and Xie, Z. (2001). DAMBE: software package for data analysis in molecular biology and evolution. *J. Hered.* 92, 371–373. doi: 10.1093/jhered/92.4.371

Xia, X., Xie, Z., Salemi, M., Chen, L., and Wang, Y. (2003). An index of substitution saturation and its application. *Mol. Phylogenet. Evol.* 26, 1–7. doi: 10.1016/S1055-7903(02)00326-3

Yacoub, H. A., Fathi, M. M., and Sadek, M. A. (2015). Using cytochrome b gene of mtDNA as a DNA barcoding marker in chicken strains. *Mitochondrial DNA* 26, 217. doi: 10.3109/19401736.2013.825771

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.