Vavilovskii Zhurnal Genetiki i Selektsii
Vavilov Journal of Genetics and Breeding. 2025;29(2):259-267

doi 10.18699/vjgb-25-29

Correction of GenBank's taxonomic entry error raises a new issue regarding intergeneric relationships among salangid fishes (Osmeriformes: Salangidae)

E.S. Balakirev

A.V. Zhirmunsky National Scientific Center of Marine Biology, Far Eastern Branch of the Russian Academy of Sciences, Vladivostok, Russia
esbalakirev@mail.ru

Abstract. The GenBank database of publicly available nucleotide sequences is the largest genetic repository providing vitally important resources for downstream applications in biology and medicine. The concern raised about reliability of GenBank data necessitates monitoring of possible taxonomic entry errors. A case of mitochondrial genome (or mitogenome) misidentification for a salangid fish belonging to the genus *Neosalanx* (Osmeriformes, Salangidae) is considered in this report. The GenBank database contains four complete mitogenome sequences of *N. taihuensis* with the accession numbers JX524196, KP170510, MH348204, and MW291630. The overall mean *p*-distance for these sequences is quite high (7.01 ± 0.14 %) but becomes 29-fold lower (0.24 ± 0.05 %) after excluding the MW291630 mitogenome. An analysis of all available nucleotide sequences of salangids has shown that the observed inconsistency in the level of divergence between *N. taihuensis* mitogenomes is due to species misidentification. It has turned out that the mitogenome MW291630 available in GenBank does not belong to *N. taihuensis*, but is, in fact, a mitogenome of *N. jordani* misidentified as *N. taihuensis*. The resolved taxonomic identity of the MW291630 mitogenome, as well as an extended sample of species with investigated single-marker sequences, has raised some new issues regarding intergeneric relationships in salangid fishes. In particular, the obtained data do not support synonymization of the genus *Neosalanx* with *Protosalanx*, as was suggested in the last revision of the salangid classification. As the comparative analysis of interspecific and intergeneric divergences shows, *Protosalanx* is not an all-inclusive clade that includes all *Neosalanx* species. Instead, it consists of (at least) two evolutionary distinct lineages with the level of genetic divergence between them matching well the mean value of divergence between the other salangid genera. Further analysis using nuclear genome-wide data is required to have new insights into the evolution of salangid fishes.

Key words: *Neosalanx*; *Protosalanx*; taxonomic misidentification; mitochondrial genomes; *CytB*; single-marker sequences; genetic divergence

For citation: Balakirev E.S. Correction of GenBank's taxonomic entry error raises a new issue regarding intergeneric relationships among salangid fishes (Osmeriformes: Salangidae). *Vavilovskii Zhurnal Genetiki i Selektsii = Vavilov J Genet Breed*. 2025;29(2):259-267. doi 10.18699/vjgb-25-29

Funding. This study received budget support within the framework of the Research Study entitled "World Ocean biodiversity: taxonomy, barcoding, phylogenetics, reproductive and evolutionary biology, biogeography" (State registry no. 121082600036-9, Ministry of Science and Higher Education of the Russian Federation) conducted by A.V. Zhirmunsky National Scientific Center of Marine Biology, Far Eastern Branch, Russian Academy of Sciences, Vladivostok, Russia.

Acknowledgements. The author greatly appreciates the encouragement and help from Elena I. Balakireva, Alla G. Oleinik, and Larisa G. Erofeeva. Evgeny P. Shvetsov (Zhirmunsky National Scientific Center of Marine Biology, FEB, RAS, Vladivostok, Russia) kindly proofread the manuscript.

Исправление таксономической ошибки в базе данных GenBank поднимает новый вопрос относительно межродовых отношений у саланксовых рыб (Osmeriformes: Salangidae)

Е.С. Балакирев

Национальный научный центр морской биологии им. А.В. Жирмунского Дальневосточного отделения Российской академии наук,
Владивосток, Россия
esbalakirev@mail.ru

Аннотация. Генетическая база данных GenBank является крупнейшим генетическим хранилищем, предоставляющим жизненно важные ресурсы для последующего применения в биологии и медицине. Высказанная обеспокоенность надежностью GenBank обусловливает необходимость мониторинга возможных таксономических ошибок в записях этой базы данных. Здесь мы сообщаем о случае ошибочной идентификации митохондриального генома (или митогенома) у саланксовой рыбы, принадлежащей к роду *Neosalanx* (Osmeriformes, Salangidae). База данных GenBank содержит четыре полные последовательности митогенома *N. taihuensis* с номерами доступа JX524196, KP170510, MH348204 и MW291630. Средняя *p*-дистанция между этими последовательностями довольно велика (7.01 ± 0.14 %), но становится в 29 раз меньше (0.24 ± 0.05 %) после исключения митогенома MW291630. Анализ всех доступных нуклеотидных последовательностей салангид показал, что наблюдаемое несоответствие в уровне дивергенции между митогеномами *N. taihuensis* обусловлено ошибочной идентификацией видов. Оказалось, что митогеном MW291630 не принадлежит *N. taihuensis*, а в действительности представляет митогеном *N. jordani*, ошибочно идентифицированный как *N. taihuensis*. Установленная таксономическая идентичность митогенома MW291630, а также расширенная выборка видов с исследованными маркерными последовательностями выявили некоторые новые аспекты межродовых отношений у саланксовых рыб. В частности, полученные данные не подтверждают синонимизацию рода *Neosalanx* с *Protosalanx*, как это было предложено в последней ревизии классификации салангид. Как показывает настоящий анализ, *Protosalanx* не является кладой, включающей все виды *Neosalanx*. Напротив, эта клада состоит по крайней мере из двух эволюционно разных линий, уровень генетической дивергенции между которыми соответствует межродовым значениям дивергенции у салангид. Необходим дальнейший анализ с использованием полных ядерных геномов для выяснения эволюции саланксовых рыб.

Ключевые слова: *Neosalanx*; *Protosalanx*; таксономические ошибки идентификации; митохондриальные геномы; *CytB*; маркерные последовательности; генетическая дивергенция

Introduction

The value and reliability of the GenBank database (Sayers et al., 2023) depends on the accuracy of species identification of biological samples, which is quite often not provided when based solely on morphology with an insufficient number of diagnostic characters. Species identification errors have been increasingly referred to as a serious challenge limiting the utility and reliability of public databases. In fact, for organisms such as fungi, which are notoriously difficult to distinguish, up to 20 % (Nilsson et al., 2006) or even 30 % (Hofstetter et al., 2019) of DNA sequence records in GenBank may have erroneous lineage designations. Multiple taxonomic misidentifications were reported for nuclear genome-sequenced strains of medically important lower eukaryotes (e. g., Houbraken et al., 2021), for single-marker sequences of many fishes (e. g., Li et al., 2018), and for complete mitogenomes of many higher eukaryotes, including bivalve mollusks (Salvi et al., 2021; Cunha et al., 2022), ticks (Mohamed et al., 2022), insects (Ožana et al., 2022; Kim et al., 2023), parasitic nematodes (Nielsen et al., 2014), fishes (Cheng et al., 2012; Balakirev et al., 2017, 2024; Oleinik et al., 2019; Sangster, Luksenburg, 2021a; Teske, 2021), amphibians (Mulder et al., 2016), reptiles (Simonov et al., 2018), birds (Sangster, Luksenburg, 2021b), and placental mammals (Botero-Castro et al., 2016).

A taxonomic misidentification causes discordance between the species name and the nucleotide sequence, thus, compromising downstream inferences. Consequently, it is urgently important to disclose such problematic sequences and report them as fast as possible after their deposition in GenBank in order to prevent propagation of incorrect biological information among databases and subsequent publications (e. g., Balakirev et al., 2017, 2024; Sangster, Luksenburg, 2021b).

Here, we report a case of mitochondrial genome misidentification for a salangid fish belonging to the genus *Neosalanx* Wakiya, Takahashi, 1937 (Osmeriformes, Salangidae). Salangids are endemic to East Asia and inhabit a wide range of marine, brackish-water, and freshwater habitats in China, Vietnam, Korean Peninsula, Japan, and Russia (e. g., Roberts, 1984). These are small, neotenic fishes with early maturation, relatively high fecundity, and a life span of about one year. Species identification of salangid fishes remains a serious challenge.

The taxonomy of salangids, based on morphological, ecological, and genetic approaches, has been subject to various revisions with multiple known synonyms (Fu et al., 2005, 2012; Zhang et al., 2007; Guo et al., 2011). In particular, it was shown that *N. taihuensis* Chen, 1956, *N. tangkahkeii* (Wu, 1931), and *N. pseudotaihuensis* Zhang, 1987 are junior synonyms of *N. brevirostris* (Pellegrin, 1923) (Zhang et al., 2007; Guo et al., 2011). *Hemisalanx* Regan, 1908 was shown to be a junior synonym of *Salanx* Cuvier, 1816 (Guo et al., 2011). Somewhat close genetic relationships were also found (Zhang et al., 2007) between *Protosalanx chinensis* (Basilewsky, 1855), *N. anderssoni* (Rendahl, 1923), *N. argentea* (Lin, 1932), and *N. tangkahkeii*. Based on the morphological characters, ecological preferences, and genetic data (mitochondrial *CytB* gene), Zhang et al. (2007) identified a group of species within the genus *Neosalanx*, including *N. reganius* Wakiya, Takahashi, 1937, *N. jordani* Wakiya, Takahashi, 1937, *N. oligodontis* Chen, 1956, and *Neosalanx* sp., which they proposed to treat as a separate new undescribed genus "*Microsalanx*". Zhang et al. (2007) and Guo et al. (2011) assumed that *N. anderssoni* may also belong to the genus *Protosalanx* Regan, 1908. Using extensive morphological analysis and also genetic markers such as mitochondrial (*CytB*) and seven nuclear genes, Fu et al. (2012) suggested that the genus *Neosalanx* should be considered a junior synonym of *Protosalanx*. These authors also found a distant relationship between *Salangichthys ishikawae* Wakiya, Takahashi, 1913 and *S. microdon* Bleeker, 1860, which proved that the two species belong to different genera: *Salangichthys* Bleeker, 1859 and the newly established *Neosalangichthys* Fu, Li, Xia, Lei, 2012 including a single species, *N. ishikawae*. Fu et al. (2012) found that the genera *Leucosoma* Gray, 1831 and *Salanx* differ significantly in genetic and morphological diagnostic characters and are, therefore, not synonymous.

Yang et al. (2020) deposited a complete mitogenome of the salangid *N. taihuensis* to GenBank under the accession no. MW291630 (taxonomy ID NCBI:txid240825), providing the forth mitogenome for this species in addition to the already available ones: JX524196, KP170510, and MH348204. An analysis of the new *N. taihuensis* MW291630 mitogenome in comparison with all other available mitogenome sequences, as well as the use of single-marker sequences of salangid fishes, has shown that this mitogenome sequence does not belong to *N. taihuensis*. We found that the specimen investigated by Yang et al. (2020) was erroneously identified as *N. taihuensis* and actually represents *N. jordani*. Therefore, the aim of the present study was to document this GenBank entry error and use the correctly identified MW291630 mitogenome, as well as an extended sample of single-marker sequences, to clarify some challenging issues regarding intergeneric relationships among salangid fishes.

Materials and methods

Mitochondrial genomes and single-marker sequences. A total of 13 complete mitogenome sequences from fishes of the family Salangidae Bleeker, 1859 were accessed from the Genetic Sequence Data Bank (the National Center for Biotechnology Information; NCBI-GenBank Flat File Release 260.0, April 15, 2024). The outgroup species, including *Plecoglossus altivelis* (Temminck, Schlegel, 1846) (family Plecoglossidae Bleeker, 1859) and *Retropinna retropinna* (Richardson, 1848) (family Retropinnidae Gill, 1862), were selected based on the previous molecular evidence of their close relationship to the family Salangidae (Fu et al., 2005; Zhang et al., 2007; Guo et al., 2011) and on a screening of nucleotide sequences available in GenBank using the basic local alignment search tool (BLAST) procedure (Altschul et al., 1990). Additionally, we also analyzed 406 mitochondrial single-marker sequences, including *12S rRNA*, *16S rRNA*, *ND1*, *COI*, and *CytB* published in previous studies on salangids (see Supplementary Table S1 for accession numbers and references)[1].

DNA sequence analysis. Previously, we described the DNA sequence analysis in detail elsewhere (e. g., Balakirev et al., 2017, 2020; Balakirev, 2022). The main steps are summarized in brief below. The nucleotide sequences were aligned using the software MUSCLE (Edgar, 2004). The programs DnaSP v. 6 (Rozas et al., 2017) and MEGA v. 11 (Tamura et al., 2021) were used for intra- and interspecific analysis of polymorphism and divergence based on uncorrected *p*-distance (Kartavtsev, 2011; Collins et al., 2012). Phylogenetic reconstructions were inferred from an analysis of complete mitogenomes by the maximum likelihood methods available in IQ-TREE v. 2 (Nguyen et al., 2015). The trees were constructed using complete mitogenomes or mitochondrial single-marker sequences only (*12S rRNA*, *16S rRNA*, *ND1*, *COI*, and *CytB*). For all reconstructions, the best-fit model of nucleotide substitution was chosen with the Akaike Information Criterion and the Bayesian Information Criterion in MEGA and IQ-TREE. The ultrafast maximum likelihood bootstrap analysis (Hoang et al., 2018) consisted of 10,000 replicates.

Results and discussion

Variability and divergence of salangid mitogenomes

Figure 1 displays a maximum likelihood tree of complete mitogenome sequences for the salangid species including representatives of the valid genera *Salanx*, *Leucosoma*, *Salangichthys*, *Protosalanx*, and *Neosalanx*. The tree shows the *N. taihuensis* (with synonyms) specimens present in two significantly diverged clusters (Lineage 1 and Lineage 2; Fig. 1) with the overall mean distance equal to 7.01 ± 0.14 %. The *N. taihuensis* mitogenome sequences from Lineage 1 (JX524196, KP170510, and MH348204) were very similar to each other (with an average *p*-distance of 0.24 ± 0.05 %), thus, demonstrating a typical level of intraspecific nucleotide diversity in fishes (e. g., Kartavtsev et al., 2016; Li et al., 2018). Lineage 1, except for *N. taihuensis*, also included *P. chinensis* and *N. anderssoni*. The overall mean distance between the species from Lineage 1 (using a single randomly picked sequence per species) was 7.70 ± 0.17 % with pairwise *p*-distances varying from 4.82 ± 0.17 % between *P. chinensis* and *N. anderssoni* to 9.21 ± 0.22 % between *P. chinensis* and *N. taihuensis*, which matched well the known interspecific nucleotide diversity in fishes (e. g., Kartavtsev et al., 2016; Li et al., 2018). Lineage 2 (Fig. 1) included the *N. taihuensis* MW291630 mitogenome only, which demonstrated a high level of divergence (14.08 ± 0.27 %) with the representative sequence of the *N. taihuensis* mitogenome from Lineage 1. With the use of all mitogenomes for the species from Lineage 1 for comparison (*P. chinensis*, *N. anderssoni*, and *N. taihuensis*; Fig. 1), the difference between Lineages 1 and 2 still remained markedly higher (13.78 ± 0.24 %) than the overall mean distance (7.70 ± 0.17 %) estimated for Lineage 1.

We found the diagnostic 15-bp deletion that occurs within the *ND5* gene (at coordinates 79–93, Supplementary Fig. S2) and the 1-bp and 2-bp diagnostic deletions that occur within the non-coding (control) region (at coordinates 534, 963, 1051–1052, and 1071; Supplementary Fig. S3). These are shared by the *P. chinensis*, *N. anderssoni*, and *N. taihuensis* mitogenomes (Lineage 1) and distinguish them clearly from the *N. taihuensis* MW291630 (Lineage 2) and the rest of the salangid mitogenomes. The 15-bp deletion within the *ND5* gene is the only sequence length variability detected for the protein-coding genes in the mitogenomes of salangid fishes. Taking into account the high phylogenetic informativeness of gaps (e. g., Giribet, Wheeler, 1999), these diagnostic deletions provide robust evidence for the close relationships of the species belonging to Lineage 1 and their distinct difference from Lineage 2.

To scale the value of full mitogenome divergence between Lineages 1 and 2, we estimated the average level of divergence based on the representative genera including *Protosalanx*, *Salanx*, *Leucosoma*, and *Salangichthys*. To be conservative, we excluded *N. anderssoni* and *N. taihuensis* (with synonyms) in order to prevent underestimation of *p*-distance values due

---

[1] Supplementary Table S1 and Figs. S1–S3 are available at: https://doi.org/10.5281/zenodo.13455533

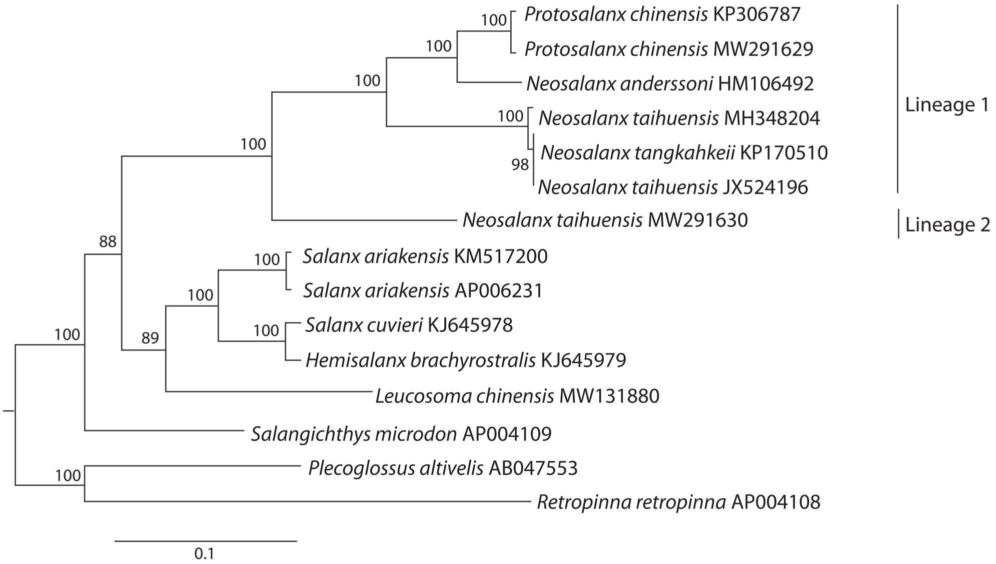

Fig. 1. Maximum likelihood tree inferred from an analysis of the complete mitochondrial genomes for fishes of the family Salangidae.

The TIM2+F+I+G4 model was used to construct the tree. The numerals at the nodes are bootstrap probability (percentage) values based on 10,000 replicates (values lower than 75 % are omitted). The tree includes all salangid mitogenomes available in GenBank except the three recombinant sequences of *Protosalanx chinensis* under the accession nos. HM106494, MH330683, and KJ499917 (Balakirev, 2022). The synonymous species names *N. taihuensis* or *N. tangkahkeii* were used for the originally published KP170510, MW291630, JX524196, and MH348204 mitogenomes. To avoid any confusion, we leave the names as they were originally assigned for the salangid species considered in this paper.

Table 1. Pairwise *p*-distances between salangid genera based on complete mitogenomes

Genus	*Salanx*	*Leucosoma*	*Salangichthys*	*Protosalanx*
Salanx		0.0026	0.0026	0.0028
Leucosoma	0.1251		0.0027	0.0029
Salangichthys	0.1342	0.1461		0.0029
Protosalanx	0.1507	0.1704	0.1658	

Note. The salangid genera *Salanx, Leucosoma*, *Salangichthys*, and *Protosalanx* are represented by the following species: *Salanx ariakensis* Kishinouye, 1902 (AP006231), *Leucosoma chinensis* (Osbeck, 1765) (MW131880), *Salangichthys microdon* (AP004109), and *Protosalanx chinensis* (KP306787). The *p*-distances are below the diagonal line. The standard errors, obtained with 10,000 bootstrap replications, are above the diagonal line.

to possible congeneric relationships of these species (Zhang et al., 2007; Guo et al., 2011). We also excluded the MW291630 mitogenome sequence with uncertain identity. The obtained overall mean *p*-distance for all available genera of salangid fishes was 14.87 ± 0.21 % with pairwise *p*-distances varying from 12.51 ± 0.26 % between *Leucosoma* and *Salanx* to 17.04 ± 0.29 % between *Leucosoma* and *Protosalanx* (Table 1), which was close to the value of divergence between Lineages 1 and 2 (13.78 ± 0.24 %).

It is worth noting that in pairwise comparisons the divergence between Lineages 1 and 2 (13.78 ± 0.24 %) was not markedly different from the divergence between *Leucosoma* and *Salangichthys*, or was even slightly higher than the divergence between *Leucosoma* and *Salanx*, as well as between *Salangichthys* and *Salanx* (Table 1). Thus, the pairwise comparisons showed that the mitogenome divergence between Lineage 1 and Lineage 2 matched well the intergeneric level of divergence in salangid fishes. The interlineage distance matched also the average value of divergence between different genera reported for the single-marker sequences or complete mitogenomes in other groups of fishes (e. g., Kartavtsev et al., 2016; Li et al., 2018; Balakirev et al., 2020).

Identification of the MW291630 mitogenome

According to a taxonomic hypothesis based on genetic data, *N. taihuensis*, *N. tangkahkeii*, *N. pseudotaihuensis*, and *N. brevirostris* are synonyms (Zhang et al., 2007; Guo et al., 2011). Consequently, the genus *Neosalanx* is represented in GenBank by only two species, *N. taihuensis* (with synonyms) and *N. anderssoni* (HM106492; Supplementary Table S1), which makes the identification of the problematic complete mitogenome MW291630 impossible. However, the GenBank database contains at least five more *Neosalanx* species, representing the full taxonomic diversity known for the genus *Neosalanx*, that were investigated using mitochondrial single-marker sequences: *N. argentea*, *N. jordani*, *N. oligodontis*,

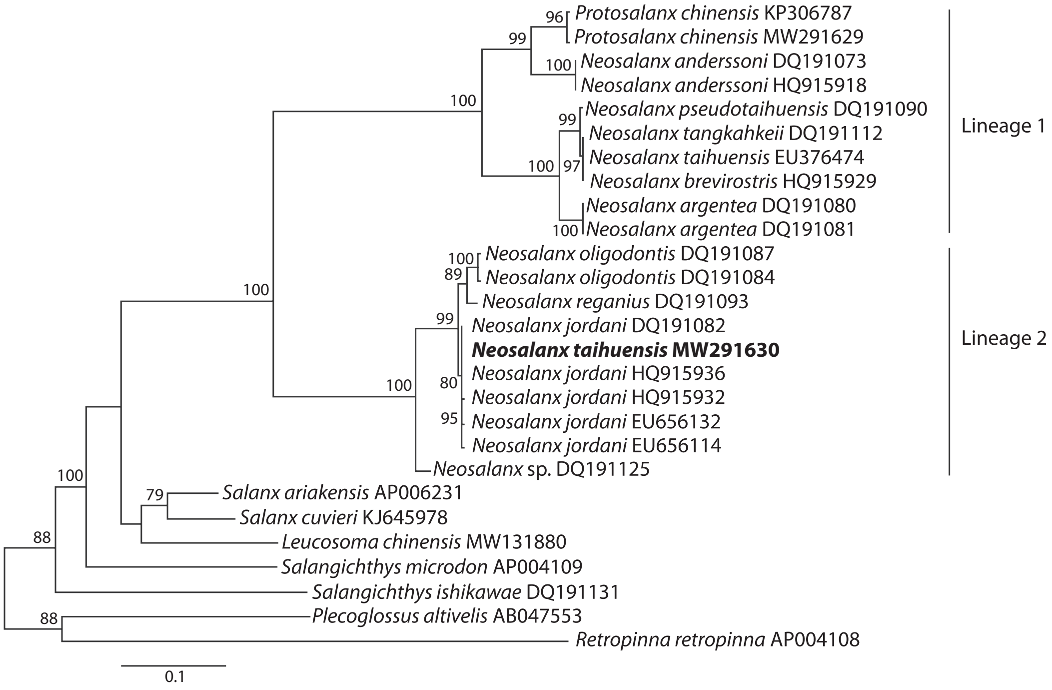

Fig. 2. Maximum likelihood tree for the members of the family Salangidae based on the *CytB* gene sequences.

The Tamura-Nei + gamma (TN93+G) model was used to infer the tree. The *N. jordani CytB* sequences are represented by three datasets investigated by Fu et al. (2012) (HQ915932 and HQ915936), Zhang et al. (2007) (DQ191082), and Zhao et al. (2010) (EU656114 and EU656132). The *N. taihuensis* MW291630 mitogenome is highlighted in bold. For tree reconstruction, we used only some representative samples from larger datasets (a full list of the *CytB* sequences is provided in Supplementary Table S1). For other comments, see Figure 1.

N. reganius, and *Neosalanx* sp. (the names of the species are listed as they were identified by the authors who submitted the respective nucleotide sequences to GenBank). The nucleotide sequences obtained for these species can be used to resolve the observed inconsistency detected for the *N. taihuensis* complete mitogenomes and to identify the taxonomically problematic MW291630 mitogenome.

We analyzed the GenBank mitochondrial single-marker sequences that are most frequently used in taxonomic and phylogenetic reconstructions of salangid fishes, including *12S rRNA*, *16S rRNA*, *ND1*, *COI*, and *CytB*. A preliminary analysis revealed that among the single-marker sequences, only *CytB* demonstrated noticeable divergence values. The other markers provided much lower resolution but were still not contradictory to the *CytB* data (see, e. g., the maximum likelihood tree based on the *COI* gene; Supplementary Fig. S1). Consequently, further analysis was based on the *CytB* gene only.

Figure 2 illustrates the maximum likelihood tree based on the *CytB* gene for *N. taihuensis* and other members of the family Salangidae representing almost the entire taxonomic diversity of the genus *Neosalanx*. There were two significantly different clusters that included the species name *N. taihuensis*. These clusters corresponded to Lineages 1 and 2 identified on the basis of mitogenome sequences (Fig. 1). The overall mean *p*-distance for Lineage 1 was 9.13 ± 0.66 % using a single randomly picked sequence per species (with pairwise *p*-distances varying from 3.33 ± 0.54 % between *N. taihuensis* and *N. argentea* to 12.09 ± 0.99 % between *N. taihuensis* and *N. anderssoni*). Lineage 1 included *P. chinensis* and part of the *Neosalanx* species (*N. anderssoni*, *N. taihuensis*, and *N. argentea*) that Fu et al. (2012) had combined with other *Neosalanx* species and synonymized with the genus *Protosalanx* (see Introduction). The second cluster contained *N. oligodontis*, *N. reganius*, *N. jordani*, *Neosalanx* sp., and the *CytB* portion of the MW291630 mitogenome with an overall mean *p*-distance of 4.69 ± 0.45 % and pairwise *p*-distances varying from 2.10 ± 0.43 % between *N. oligodontis* and *N. reganius* to 7.01 ± 0.76 % between *N. reganius* and *Neosalanx* sp. This cluster included a group of species that were placed in the genus "*Microsalanx*" erected by Zhang et al. (2007). The mean *p*-distance between the clusters (Lineage 1 and Lineage 2, Fig. 2) was 19.03 ± 1.06 % with a single randomly picked sequence per species or 18.97 ± 1.09 % with all 146 sequences available for Lineages 1 and 2 (Supplementary Table S1). This value fit well into the range of intergeneric divergences of fishes (e. g., Kartavtsev et al., 2016; Li et al., 2018; Balakirev et al., 2020).

An analysis of pairwise *p*-distances showed a surprisingly high level of sequence divergence (18.89 ± 1.16 %) (Table 2) between the GenBank *CytB* sequences of *N. taihuensis*, including 70 isolates obtained from different localities by various authors (Zhang et al., 2007; Zhao et al., 2008; see also Supplementary Table S1) with low intraspecific variability (0.44 ± 0.09 %), and the respective *CytB* portion of the MW291630 mitogenome. The result was consistent with the data obtained for the complete mitogenomes (see above), which showed a substantial difference between the MW291630

Table 2. Pairwise *p*-distances between the *CytB* sequences for *Neosalanx* species

	1	2	3	4	5	6	7	8
1		0.0003	0.0041	0.0044	0.0075	0.0116	0.0119	0.0120
2	0.0012		0.0040	0.0044	0.0075	0.0115	0.0119	0.0120
3	0.0189	0.0196		0.0042	0.0078	0.0114	0.0117	0.0119
4	0.0219	0.0229	0.0198		0.0077	0.0115	0.0117	0.0120
5	0.0657	0.0666	0.0686	0.0701		0.0112	0.0114	0.0115
6	0.1889	0.1890	0.1837	0.1853	0.1848		0.0052	0.0099
7	0.1906	0.1909	0.1855	0.1836	0.1862	0.0332		0.0097
8	0.2004	0.2003	0.1964	0.1986	0.1841	0.1215	0.1172	

Note. All available *CytB* sequences for each included species were used for this analysis (the number of sequences is in parentheses). 1: MW291630 (1); 2: *N. jordani* (25); 3: *N. oligodontis* (7); 4: *N. reganius* (1); 5: *Neosalanx* sp. (1); 6: *N. taihuensis* (70); 7: *N. argentea* (2); and 8: *N. anderssoni* (10). For other comments, see Table 1 and Figure 2.

mitogenome and the other *N. taihuensis* (with synonyms) mitogenomes, JX524196, KP170510, and MH348204, available in GenBank.

The *CytB* portion of the *N. taihuensis* MW291630 mitogenome demonstrated very close affinity to the *N. jordani* single-marker sequences obtained from different localities by various authors (25 isolates; Zhang et al., 2007; Zhao et al., 2010; Fu et al., 2012) with low intraspecific variability (0.24 ± 0.05 %; see also Fig. 2). The *p*-distance between the *CytB* gene of the MW291630 mitogenome and the GenBank *CytB* sequences of *N. jordani* was surprisingly low (0.12 ± 0.03 %; Table 2); it fit well into the range of intraspecific variability in fishes (e.g., Kartavtsev et al., 2016; Li et al., 2018). The maximum likelihood tree based on the *COI* gene showed a similar topology (Supplementary Fig. S1).

Three species, *N. oligodontis*, *N. reganius*, and *Neosalanx* sp., demonstrated more pronounced differences from the MW291630 mitogenome than *N. jordani* (Table 2, Fig. 2). Zhang et al. (2007) suggested that *N. reganius* and *N. oligodontis* could be considered as subspecies of *N. jordani*. Indeed, *N. jordani*, *N. oligodontis*, *N. reganius*, *Neosalanx* sp., and the *CytB* portion of the MW291630 mitogenome were all in the same cluster (Fig. 2) with an overall mean *p*-distance of 4.69 ± 0.45 %, which suggested close relationships for these salangid species.

Thus, the single-marker sequences clearly showed that the GenBank accession no. MW291630 represents the mitogenome of *N. jordani* mistaken for *N. taihuensis*. The observed inconsistency in the level of divergence between the *N. taihuensis* mitogenomes (see above) is due to incorrect species identification. The *CytB* analysis of within- and between lineage variability confirmed the data obtained with complete mitogenomes.

MtDNA data indicates a generic level of divergence between Lineage 1 and Lineage 2

The close relationships between *N. jordani*, *N. oligodontis*, *N. reganius*, and *Neosalanx* sp. had been reported previously, as well as the significant difference of this group from other *Neosalanx* and *Protosalanx* species including *N. taihuensis*, *N. argentea*, *N. anderssoni*, and *P. chinensis* (Fig. 2, Table 2) (Zhang et al., 2007; Guo et al., 2011). Based on integrative data, Zhang et al. (2007) erected the genus "*Microsalanx*" (see Introduction). The results of the present analysis do not contradict this hypothesis. Both the complete mitogenomes (Fig. 1) and the single-marker sequences (Fig. 2) clearly demonstrated two significantly diverged clusters (Lineage 1 and Lineage 2). The interlineage divergence based on the *CytB* gene was high (18.97 ± 1.09 %). It was significantly higher than the average divergences within each of the lineages: the overall mean distances for Lineage 1 and Lineage 2 separately were 9.13 ± 0.66 and 4.69 ± 0.45 %, respectively.

To evaluate the scale of divergence between Lineages 1 and 2, we estimated the average level of divergence based on the *CytB* gene for the salangid genera available in GenBank including *Protosalanx*, *Salanx*, *Leucosoma*, *Neosalangichthys*,

Table 3. Pairwise *p*-distances between salangid genera based on the *CytB* gene

Genus	*Protosalanx*	*Salanx*	*Leucosoma*	*Neosalangichthys*	*Salangichthys*
Protosalanx		0.0118	0.0123	0.0124	0.0125
Salanx	0.1876		0.0093	0.0106	0.0108
Leucosoma	0.2068	0.1174		0.0103	0.0113
Neosalangichthys	0.2077	0.1472	0.1411		0.0109
Salangichthys	0.2147	0.1490	0.1648	0.1613	

Note. The salangid genera *Protosalanx*, *Salanx*, *Leucosoma*, *Neosalangichthys*, and *Salangichthys* are represented by the following species: *Protosalanx chinensis* (KP306787), *Salanx ariakensis* (AP006231), *Leucosoma chinensis* (MW131880), *Neosalangichthys ishikawae* (Wakiya, Takahashi, 1913) (DQ191127), and *Salangichthys microdon* (AP004109). For other comments, see Table 1.

and *Salangichthys* (Table 3). The obtained overall mean *p*-distance was 16.98 ± 0.78 % (with pairwise *p*-distances varying from 11.74 ± 0.93 % between *Leucosoma* and *Salanx* to 21.47 ± 1.25 % between and *Protosalanx* and *Salangichthys*; Table 3), which was close to the value of divergence between Lineages 1 and 2 (18.97 ± 1.09 %) based on the multiple *CytB* gene sequences (see above). The divergence between Lineages 1 and 2 (18.97 ± 1.09 %) was not markedly different from that between *Protosalanx* and *Salanx* or it was even higher than the *p*-distances in pairwise comparisons of *Leucosoma* vs. *Salanx*, *Neosalangihthys*, and *Salangichthys*; *Salanx* vs. *Neosalangihthys* and *Salangichthys*; and *Neosalangihthys* vs. *Salangichthys* (Table 3).

Thus, an analysis of the multiple *CytB* sequences encompassing the full diversity of salangid fishes showed a high level of divergence between Lineage 1 and Lineage 2 (18.97 ± 1.09 %), which fit into the range of intergeneric distances reported for salangids (see above) and other fishes (see the references above). Lineage 1 included a group of species (*P. chinensis*, *N. anderssoni*, *N. taihuensis*, and *N. argentea*; Fig. 2) comprising a part of the reorganized genus *Protosalanx* (Fu et al., 2012). The group of species from Lineage 1 was previously divided in two sub-groups (“primitive lineages”) (Zhang et al., 2007; Guo et al., 2011). Indeed, the pairwise *p*-distances for Lineage 1 varied within a relatively wide range from 3.32 ± 0.52 % between *N. taihuensis* and *N. argentea* to 12.15 ± 0.99 % between *N. taihuensis* and *N. anderssoni* (Table 2). However, the overall mean sub-group divergence (*P. chinensis* + *N. anderssoni* vs. *N. taihuensis* + *N. argentea*) within Lineage 1 was still markedly lower (11.42 ± 0.87 %) than the divergence between Lineages 1 and 2 (18.97 ± 1.09 %). Thus, unlike Zhang et al. (2007) and Guo et al. (2011), we did not find sufficient grounds to split Lineage 1 into two sub-groups and consider it a single evolutionary lineage representing the genus *Protosalanx*. This conclusion was supported by the diagnostic deletions detected within the *ND5* gene and the control region in the salangids’ mitogenomes (see the “Variability and divergence of salangid mitogenomes” section above). Nevertheless, the relationships between the “primitive lineages” *P. chinensis* + *N. anderssoni* and *N. taihuensis* + *N. argentea* need to be further clarified using a more representative array of genetic markers (see below).

Lineage 2 contained a group of species (*N. oligodontis*, *N. reganius*, *N. jordani*, and *Neosalanx* sp.) placed in the genus “*Microsalanx*” by Zhang et al. (2007). This subdivision was reasonable (see Introduction) to distinguish this group of species from the rest of the *Neosalanx* species. However, the transfer of *N. taihuensis* (with synonyms), *N. anderssoni*, and *N. argentea* to the genus *Protosalanx*, as suggested earlier (Zhang et al., 2007; Guo et al., 2011; Fu et al., 2012) and supported by our data (Figs. 1 and 2, Tables 1 and 2), gives reason to abolish the genus name “*Microsalanx*” (at least until the generic heterogeneity is proven for Lineage 1; see above). Consequently, the original genus name *Neosalanx* is appropriate for the salangid species *N. oligodontis*, *N. reganius*, *N. jordani*, and *Neosalanx* sp. comprising Lineage 2 (Figs. 1 and 2).

Thus, in contrast to Fu et al. (2012), our analysis based on complete mitogenomes and mtDNA single-marker sequences, as well as the analysis of Zhang et al. (2007) based on morphological, ecological, and genetic data, did not support the synonymization of all *Neosalanx* species with *Protosalanx*. The data clearly show two substantially diverged evolutionary lineages (Figs. 1 and 2): (1) *P. chinensis*, *N. anderssoni*, *N. taihuensis* (with synonyms), and *N. argentea* representing the genus *Protosalanx* and (2) *N. oligodontis*, *N. reganius*, *N. jordani*, and *Neosalanx* sp. representing the genus *Neosalanx*.

For phylogenetic analysis of salangid fishes, Fu et al. (2012) used a concatenated multigene dataset including the mitochondrial *CytB* gene and seven nuclear sequences (*28S rRNA*, *RAG1*, *zic1*, *ENC1*, *RNF213*, *glyt*, and *SH3PX3*). As a result (among others), these authors (Fu et al., 2012, p. 853) discovered that “all species from the ‘*Neosalanx–Protosalanx*’ complex belong to a same genus” and considered *Neosalanx* as a junior synonym of *Protosalanx*.

Compared to mtDNA markers and complete mitogenomes, the nuclear markers (*28S rRNA*, *RAG1*, *zic1*, *ENC1*, *RNF213*, *glyt*, and *SH3PX3*), mostly used by Fu et al. (2012), demonstrated a much lower divergence between the salangid genera. For the genera *Protosalanx* (except *Neosalanx*), *Salanx*, *Leucosoma*, *Neosalangichthys*, and *Salangichthys*, the values of the overall mean distance for the nuclear markers were low and varied in a narrow range (from 1.98 ± 0.36 % for *zic1* to 3.56 ± 0.54 % for *RAG1*). The low divergence of the nuclear markers can be explained by the fact that they mostly represent highly conserved sequences developed for analyzing deep phylogenetic relationships on a scale of dozens to hundreds of millions of years, e. g., to infer phylogenetic relationships of all bony fishes, which requires analysis of genomic regions with slow rates of evolution (e. g., Betancur-R et al., 2017). These markers might be not sensitive enough for salangid fishes that experienced most speciation events around 1.1–9.9 Ma (Zhang et al., 2007). As a consequence, we suggest that the phylogenetic signal of *CytB*, also used by Fu et al. (2012), was significantly “diluted” by the effect of strongly conserved nuclear sequences. Indeed, the overall mean *p*-distance between the genera *Protosalanx* (except *Neosalanx*), *Salanx*, *Leucosoma*, *Neosalangichthys*, and *Salangichthys* was equal to 16.98 ± 0.78 % based on the *CytB* gene only (see above). However, it decreased significantly, to 2.72 ± 0.17 %, when the nuclear multigene dataset of Fu et al. (2012) was used.

Although the suggested relationships in salangid fishes seem robust, we expect them to be modified, possibly, as new genetic data become available. In particular, the mitochondrial sequences have revealed a relatively high level of divergence between two sub-groups within the genus *Protosalanx* (*P. chinensis* + *N. anderssoni* and *N. taihuensis* + *N. argentea*; Fig. 2) (see also Zhang et al., 2007; Guo et al., 2011), which may indicate a supra-species taxonomical range. Consequently, more nuclear genome-wide data are necessary to further address this and other issues concerning the taxonomic composition and the evolutionary relationships among salangid fishes.

Conclusion

Misidentified nucleotide sequences, including complete mitogenomes, are becoming increasingly frequent in GenBank, which leads to an explosive spread of incorrect biological information in subsequent scientific publications over time.

The misidentified *N. taihuensis* MW291630 mitogenome has been revealed in our study. We argue that the GenBank accession no. MW291630 actually represents the mitogenome of *N. jordani* mistaken for *N. taihuensis*. Thus, GenBank users should be aware of the above-described entry error to avoid conflicting results in their downstream evolutionary and comparative genomic studies.

The data obtained have raised a new issue regarding intergeneric relationships among salangid fishes. In contrast to the study by Fu et al. (2012), our data from the comparative analyses of interspecific and intergeneric divergences do not support the synonymization of the genus *Neosalanx* with *Protosalanx* and oppose the suggestion to consider *Neosalanx* as a junior synonym of *Protosalanx*. Genome-wide studies are needed to further clarify the evolutionary relationships of salangid fishes.

The introduction and spread of misidentified nucleotide sequences in genetic databases, which compromises downstream applications, is unlikely to be completely curbed. However, some appropriate steps can be undertaken (see, e. g., Balakirev et al., 2017, 2024; Sangster, Luksenburg, 2021b) to minimize their massive accumulation and subsequent propagation in scientific publications, thereby increasing the reliability of findings reported in them.

References

Altschul S., Gish W., Miller W., Myers E., Lipman D. Basic local alignment search tool. *J Mol Biol*. 1990;215:403-410. doi 10.1016/S0022-2836(05)80360-2

Balakirev E.S. Recombinant mitochondrial genomes reveal recent interspecific hybridization between invasive salangid fishes. *Life*. 2022;5:661. doi 10.3390/life12050661

Balakirev E.S., Saveliev P.A., Ayala F.J. Complete mitochondrial genomes of the Cherskii's sculpin *Cottus czerskii* and Siberian taimen *Hucho taimen* reveal GenBank entry errors: incorrect species identification and recombinant mitochondrial genome. *Evol Bioinform Online*. 2017;13:1176934317726783. doi 10.1177/1176934317726783

Balakirev E.S., Kravchenko A.Y., Semenchenko A.A. Genetic evidence for a mixed composition of the genus *Myoxocephalus* (Cottoidei: Cottidae) necessitates generic realignment. *Genes*. 2020;11:1071. doi 10.3390/genes11091071

Balakirev E.S., Sharina S.N., Balanov A.A. Misidentified mitogenomes of two *Lycodes* species (Perciformes: Zoarcidae) in GenBank. *Russ J Genet*. 2024;60(10):1375-1382. doi 10.1134/S1022795424700911

Betancur-R R., Wiley E.O., Arratia G., Acero A., Bailly N., Miya M., Lecointre G., Ortí G. Phylogenetic classification of bony fishes. *BMC Evol Biol*. 2017;17:162. doi 10.1186/s12862-017-0958-3

Botero-Castro F., Delsuc F., Douzery E.J.P. Thrice better than once: quality control guidelines to validate new mitogenomes. *Mitochondrial DNA A DNA Mapp Seq Anal*. 2016;27(1):449-454. doi 10.3109/19401736.2014.900666

Cheng J., Ma G.-Q., Song N., Gao T.-X. Complete mitochondrial genome sequence of bighead croaker *Collichthys niveatus* (Perciformes, Sciaenidae): a mitogenomic perspective on the phylogenetic relationships of Pseudosciaeniae. *Gene*. 2012;491(2):210-223. doi 10.1016/j.gene.2011.09.020

Collins R.A., Boykin L.M., Cruickshank R.H., Armstrong K.F. Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods Ecol Evol*. 2012;3:457-465. doi 10.1111/j.2041-210x.2011.00176.x

Cunha R.L., Nicastro K.R., Zardi G.I., Madeira C., McQuaid C.D., Cox C.J., Castilho R. Comparative mitogenomic analyses and gene rearrangements reject the alleged polyphyly of a bivalve genus. *PeerJ*. 2022;10:e13953. doi 10.7717/peerj.13953

Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792-1797. doi 10.1093/nar/gkh340

Fu C., Luo J., Wu J., López J.A., Zhong Y., Lei G., Chen J. Phylogenetic relationships of salangid fishes (Osmeridae, Salanginae) with comments on phylogenetic placement of the salangids based on mitochondrial DNA sequences. *Mol Phylogenet Evol*. 2005;35:76-84. doi 10.1016/j.ympev.2004.11.024

Fu C., Guo L., Xia R., Li J., Lei G. A multilocus phylogeny of Asian noodlefishes Salangidae (Teleostei: Osmeriformes) with a revised classification of the family. *Mol Phylogenet Evol*. 2012;62(3):848-855. doi 10.1016/j.ympev.2011.11.031

Giribet G., Wheeler W.C. On gaps. *Mol Phylogenet Evol*. 1999;13(1):132-143. doi 10.1006/mpev.1999.0643

Guo L., Li J., Wang Z., Fu C. Phylogenetic relationships of noodlefishes (Osmeriformes: Salangidae) based on four mitochondrial genes. *Acta Hydrobiol*. 2011;35:449-459. doi 10.3724/SP.J.1035.2011.00449

Hoang D.T., Chernomor O., von Haeseler A., Minh B.Q., Vinh L.S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 2018;35:518-522. doi 10.1093/molbev/msx281

Hofstetter V., Buyck B., Eyssartier G., Schnee S., Gindro K. The unbearable lightness of sequenced-based identification. *Fungal Divers*. 2019;96:243-284. doi 10.1007/s13225-019-00428-3

Houbraken J., Visagie C.M., Frisvad J.C. Recommendations to prevent taxonomic misidentification of genome-sequenced fungal strains. *Microbiol Resour Ann*. 2021;10:e01074-20. doi 10.1128/MRA.01074-20

Kartavtsev Y.P. Sequence divergence at mitochondrial genes in animals: applicability of DNA data in genetics of speciation and molecular phylogenetics. *Mar Genomics*. 2011;4(2):71-81. doi 10.1016/j.margen.2011.02.002

Kartavtsev Y.P., Rozhkovan K.V., Masalkova N.A. Phylogeny based on two mtDNA genes (*Co-1*, *Cyt-B*) among sculpins (Scorpaeniformes, Cottidae) and some other scorpionfish in the Russian Far East. *Mitochondrial DNA A DNA Mapp Seq Anal*. 2016;27(3):2225-2240. doi 10.3109/19401736.2014.984164

Kim D.E., Kim P., Lee H., Kim N.H., Kim D., Lee M.J., Ban Y.G., Jang B., Park J. Comprehensive analysis of the complete mitochondrial genome of *Melanoplus differentialis* (Acrididae: Melanoplinae) captured in Korea. *Entomol Res*. 2023;53:66-81. doi 10.1111/1748-5967.12633

Li X., Shen X., Chen X., Xiang D., Murphy R.W., Shen Y. Detection of potential problematic *Cytb* gene sequences of fishes in GenBank. *Front Genet*. 2018;9:30. doi 10.3389/fgene.2018.00030

Mohamed W.M.A., Moustafa M.A.M., Kelava S., Barker D., Matsuno K., Nonaka N., Shao R., Mans B.J., Barker S.C., Nakao R. Reconstruction of mitochondrial genomes from raw sequencing data provides insights on the phylogeny of *Ixodes* ticks and cautions for species misidentification. *Ticks Tick Borne Dis*. 2022;13(1):101832. doi 10.1016/j.ttbdis.2021.101832

Mulder K.P., Lourenço A., Carneiro M., Velo-Antón G. The complete mitochondrial genome of *Salamandra salamandra* (Amphibia: Urodela: Salamandridae). *Mitochondrial DNA Part B*. 2016;1:880-882. doi 10.1080/23802359.2016.1253042

Nguyen L.T., Schmidt H.A., von Haeseler A., Minh B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. *Mol Biol Evol*. 2015;32:268-274. doi 10.1093/molbev/msu300

Nielsen M.K., Wang J., Davis R., Bellaw J.L., Lyons E.T., Lear T.L., Goday C. *Parascaris univalens* – a victim of large-scale misidentification? *Parasitol Res*. 2014;113:4485-4490. doi 10.1007/s00436-014-4135-y

Nilsson R.H., Ryberg M., Kristiansson E., Abarenkov K., Larsson K.H., Koljalg U. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. *PLoS One*. 2006;1(1):e59. doi 10.1371/journal.pone.0000059

Oleinik A.G., Skurikhina L.A., Kukhlevsky A.D. Clarification of taxonomic assignment of smelt complete mitochondrial genome:

GenBank accession number KP281293.1 (NC_026566.1). *Mitochondrial DNA Part B*. 2019;4:1696-1697. doi 10.1080/23802359.2019.1607578

Ožana S., Dolný A., Pánek T. Nuclear copies of mitochondrial DNA as a potential problem for phylogenetic and population genetic studies of Odonata. *Syst Entomol*. 2022;47:591-602. doi 10.1111/syen.12550

Roberts T.R. Skeletal anatomy and classification of the neotenic Asian Salmoniform superfamily Salangoidea (icefishes or noodlefishes). *Proc Califor Acad Sci*. 1984;43:179-220

Rozas J., Ferrer-Mata A., Sánchez-DelBarrio J.C., Guirao-Rico S., Librado P., Ramos-Onsins S.E., Sánchez-Gracia A. DnaSP 6: DNA Sequence Polymorphism analysis of large datasets. *Mol Biol Evol*. 2017;34:3299-3302. doi 10.1093/molbev/msx248

Salvi D., Berrilli E., Garzia M., Mariottini P. Yet another mitochondrial genome of the Pacific cupped oyster: the published mitogenome of *Alectryonella plicatula* (Ostreinae) is based on a misidentified *Magallana gigas* (Crassostreinae). *Front Mar Sci*. 2021;8:741455. doi 10.3389/fmars.2021.741455

Sangster G., Luksenburg J.A. The published complete mitochondrial genome of the milk shark (*Rhizoprionodon acutus*) is a misidentified Pacific spadenose shark (*Scoliodon macrorhynchos*) (Chondrichthyes: Carcharhiniformes). *Mitochondrial DNA Part B*. 2021a;6:828-830. doi 10.1080/23802359.2021.1884019

Sangster G., Luksenburg J.A. Sharp increase of problematic mitogenomes of birds: causes, consequences, and remedies. *Genome Biol Evol*. 2021b;13:evab210. doi 10.1093/gbe/evab210

Sayers E.W., Cavanaugh M., Clark K., Pruitt K.D., Sherry S.T., Yankie L., Karsch-Mizrachi I. GenBank 2023 update. *Nucleic Acids Res*. 2023;51:D141-D144. doi 10.1093/nar/gkac1012

Simonov E., Lisachov A., Oreshkova N., Krutovsky K.V. The mitogenome of *Elaphe bimaculata* (Reptilia: Colubridae) has never been published: a case with the complete mitochondrial genome of *E. dione*. *Acta Herpetol*. 2018;13:185-189. doi 10.13128/Acta_Herpetol-23394

Tamura K., Stecher G., Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis version 11. *Mol Biol Evol*. 2021;38:3022-3027. doi 10.1093/molbev/msab120

Teske P.R. Mitochondrial genome announcements need to consider existing short sequences from closely related species to prevent taxonomic errors. *Conserv Genet Resour*. 2021;13:359-365. doi 10.1007/s12686-021-01214-7

The National Center for Biotechnology Information. Available online: https://www.ncbi.nlm.nih.gov/ (accessed on July 29, 2024)

Yang Y., Sui Z., Liu K., Liu Y. The complete mitochondrial DNA sequence of Linyi small icefish (*Neosalanx taihuensis*). GenBank submission: 24-NOV-2020. Genbank accession number: MW291630

Zhang J., Li M., Xu M., Takita T., Wei F. Molecular phylogeny of icefish Salangidae based on complete mtDNA *cytochrome b* sequences, with comments on estuarine fish evolution. *Biol J Linn Soc*. 2007;91:325-340. doi 10.1111/j.1095-8312.2007.00785.x

Zhao L., Zhang J., Liu Z., Funk S.M., Wei F., Xu M., Li M. Complex population genetic and demographic history of the Salangid, *Neosalanx taihuensis*, based on cytochrome *b* sequences. *BMC Evol Biol*. 2008;8:201. doi 10.1186/1471-2148-8-201

Zhao L., Zhang J., Liu Z., Xu M., Li M. Population genetic structure and demographic history of *Neosalanx jordani* based on cytochrome *b* sequences. *Biodiv Sci*. 2010;18(3):251-261. doi 10.3724/SP.J.1003.2010.251

Conflict of interest. The author declares no conflicts of interest.

Received June 9, 2024. Revised August 29, 2024. Accepted October 25, 2024.