Complete mitochondrial genomes and updated divergence time of the two freshwater clupeids endemic to Lake Tanganyika (Africa) suggest intralacustrine speciation

Leona J.M. Milec (milecleona@hotmail.com)
Nord University

Maarten P.M. Vanhove
Hasselt University

Fidel Muterezi Bukinga
Centre de Recherche en Hydrobiologie Uvira

Els L.R. Keyzer
Universiteit Antwerpen

Vercus Lumami Kapepula
Université Catholique de Louvain

Pascal Masilya Mulungula
Centre de Recherche en Hydrobiologie Uvira

N'Sibula Mulimbwa
Centre de Recherche en Hydrobiologie Uvira

Catherine E. Wagner
University of Wyoming

Joost A.M. Raeymaekers
Nord University

Research Article

Keywords: Great Lakes, Clupeiformes, Radiation, Mitogenome, Time calibration, Phylogenetics

Posted Date: July 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1877994/v1

Abstract

Background

The hydro-geological history of Lake Tanganyika paints a complex image of several colonization and adaptive radiation events. The initial basin was formed around 9-12 million years ago (MYA) from the predecessor of the Malagarasi–Congo River and only 5-6 MYA, its sub-basins fused to produce the clear, deep waters of today. Next to the well-known radiations of cichlid fishes, the lake also harbours a modest clade of only two clupeid species, *Stolothrissa tanganicae* and *Limnothrissa miodon.* They are members of Pellonulini, a tribe of clupeid fishes that mostly occur in freshwater and that colonized West- and Central-Africa during a period of high sea levels during the Cenozoic. There is no consensus on the phylogenetic relationships between members of Pellonulini and the timing of the colonization of Lake Tanganyika by clupeids.

Results

We use short-read next generation sequencing of 10X Chromium libraries to sequence and assemble the full mitochondrial genomes of *S. tanganicae* and *L. miodon.* We then use Maximum Likelihood and Bayesian methods to place them into the phylogeny of Pellonulini and other clupeiforms, taking advantage of all available full mitochondrial clupeiform genomes. We identify *Potamothrissa obtusirostris* as the closest living relative of the Tanganyika sardines and establish paraphyly for *Microthrissa.* Our results indicate a relatively recent divergence of the Tanganyika sardines around 4.07 MYA [95% CI: 1.66-7.06], and from *P. obtusirostris* around 11.88 MYA [95% CI: 6.09-17.27].

Conclusions

These young estimates imply that the ancestor of the Tanganyika sardines diverged from a riverine ancestor and entered the proto-lake Tanganyika around the time of its formation from the Malagarasi–Congo river, and diverged into the two extant species at the onset of deep clearwater conditions. Our results prompt a more thorough examination of the relationships within Pellonulini, and the new mitochondrial genomes provide an important resource for the future study of this tribe*,* e.g. as a reference for species identification, genetic diversity, and macroevolutionary studies.

Background

Lake Tanganyika has experienced a turbulent geological history of lake level fluctuations, shifting shorelines and transient hydrological connections, paving the way for a complex sequence of colonisations that gave rise to a diverse freshwater fauna with a high degree of endemism (1). The lake was originally formed by lateral expansion of the western branch of the East African rift, crossing the predecessor of the Malagarasi–Congo River around 9–12 million years ago (MYA). Only 5–6 MYA its water levels rose high enough for the subbasins and the swampy areas in between to fuse into the deep clearwater lake of today (2–4). The lake has experienced large water level fluctuations since (5,6). These events are reflected in the evolutionary history of the organisms inhabiting the lake. For example, the adaptive radiations of the Tanganyika cichlid tribes happened over several stages, with some ancestral species colonizing the lake in the early stages of its formation, while others diversified later when the historical subbasin lakes fused (7,8) or following the depression of the northernmost Tanganyika basin around 7–8 MYA (9). Yet other lineages were initially thought to have colonized the lake at an even later stage, and thus established themselves in an already present adaptive radiation (10,11). Recent work, however, suggests that the cichlid radiation unfolded completely within the temporal and spatial confines of Tanganyika (9,12–14).

Next to this textbook example of adaptive radiation, Lake Tanganyika also harbours a small “flock” of two endemic clupeid species, *Stolothrissa tanganicae* and *Limnothrissa miodon*. These clupeids are members of the African clupeid tribe Pellonulini, one of the most diverse freshwater radiations of Clupeiformes with 22 species in 11 genera, most occurring either on the West coast of Africa (distribution from Senegal down to Congo/Angola), or in the Congo river system and its tributaries and lakes (15,16). The members of Pellonulini are thought to be derived from a group of sardine-like species whose ancestors originated from the Atlantic West coast of Africa during a period of high sea levels between 30–50 MYA (16–18). The exact route this radiation took through the Congo basin is unknown, and the relationships between pellonuline taxa remain inconsistent in published clupeid phylogenies (17,19–21).

The Tanganyika sardines are the fully pelagic, planktivorous, endemic *S. tanganicae* and the semi-pelagic, more opportunistic *L. miodon*, which is originally endemic to Lake Tanganyika but has also established in other lakes in Central Africa after anthropogenic introductions. Both species are important fisheries targets in Lake Tanganyika and provide food and livelihood for millions of people (22,23). The colonization and subsequent speciation of the Tanganyika sardines has only been explicitly addressed once (17), and estimated as part of larger phylogenies twice more (19,20). In these studies, estimates of their divergence time are based on minimum one and maximum three mitochondrial genes, and show substantial variation, the youngest being at 3.91 MYA and the oldest at 8 MYA with a large credibility interval (CI). Lake Tanganyika was formed 9–12 MYA, with the northern and southern subbasins forming at 7–8 MYA and 2–4 MYA respectively. The fusion of the subbasins and onset of clearwater conditions is estimated at 5–6 MYA. Keeping these estimates in mind, a divergence time of the two sardine species of 8–10 MYA would mean that the lineage leading to *S. tanganicae* and *L. miodon* started to undergo speciation soon after entering the not yet connected subbasins of the proto-lake. An older divergence time would indicate riverine speciation and subsequent colonization of the proto-lake. In contrast, a more recent divergence time would agree with intralacustrine speciation i.e. after the subbasins of the lake connected to form the deep rift lake we see today.

Robust phylogenies and estimates of divergence time between lineages are crucial to understanding the relationship between geological or hydrological events, speciation and realised biodiversity. Mitochondrial genes, especially those coding for cytochrome *c* oxidase subunit I (COI), cytochrome B (CYTB) and ribosomal RNAs (12S rRNA, 16S rRNA), have classically been used for phylogenetic studies because of their high degree of conservation, while still exhibiting enough variation to distinguish between species

(24). Single-gene datasets have limited ability to recover true phylogenetic relationships, especially in more closely related species (13,25,26) and tend to overestimate divergence time (27). Whole mitochondrial genomes can contain phylogenetic information that is lost when targeting a single gene. In contrast, the high mutation rate of certain mitochondrial regions, such as the control region or the NADH-ubiquinone oxidoreductase chain 4L (ND4L) gene, have been considered problematic in phylogenetic datasets because of saturation and ensuing homoplasy (26,27), although this does not necessarily imply lower phylogenetic information content (28). Most modern phylogenomic algorithms allow partitioned analysis of multi-locus alignments that estimate and evaluate separate evolutionary models for each partition. This enables sequences with different characteristics to be effectively combined during analyses, by modelling differences including base composition, transition-transversion biases, and heterogenous nucleotide substitution among sites (29,30). Compared to using single or a few markers, whole mitochondrial genomes have yielded higher resolution and better supported phylogenies in recent studies of fish (31,32) and other vertebrates (27,33,34), especially when investigating recently diverged or taxonomically diverse taxa (27).

In this study, we use short-read next generation sequencing (NGS) to sequence and assemble the complete mitochondrial genomes of *S. tanganicae* and *L. miodon.* We then use the new sequences, together with all available full mitochondrial clupeiform genomes, to build the first phylogeny of members of Pellonulinae to include all mitochondrial protein-coding genes (PCGs), rRNA-genes and the D-loop (control region). We revisit the phylogenetic relationships within Pellonulinae and estimate the divergence time of the Lake Tanganyika sardines with improved resolution. We discuss the results in the light of the geological history of Lake Tanganyika.

Results

New mitochondrial genomes, diversity and divergence

The mitogenome assemblies of *S. tanganicae* and *L. miodon* were 16737 bp and 16739 bp long, respectively. We annotated all 13 PCGs, 22 transfer RNA (tRNA) genes, 2 rRNA genes (total = 37 genes) as well as the control region (D-loop) in both assemblies in the typical fish and vertebrate mitochondrial gene order (35) (Fig. 1). Gene order analysis in CREx confirmed that all included clupeiforms follow the same gene order, except *Ilisha elongata*, where tRNA-Pro and tRNA-Thr appeared transposed.

Nucleotide diversity (π), calculated based on alignments of mitochondrial genes between 109 clupeiform and 6 non-clupeiform fish species, showed peaks in the beginning and end of 16S rDNA, as well as in the genes coding for ND1, ND2, ND3, ND5 and ATP synthase membrane subunit 6 (ATP6) (Fig. 2). The genes coding for COI, COII, COIII and CYTB, along with some regions of 12S and 16S rDNA, were relatively less diverse. The alignments for the D-loop, ND4L, ND4 and ND6 genes contained too many gaps to accurately calculate nucleotide diversity and were excluded from this analysis.

Analysis of pairwise genetic distance of PCGs revealed that *S. tanganicae* and *L. miodon* were among the 0.3% (PCGs) or 3% (non-coding regions) most similar species of Clupeiformes (17th or 206th most similar out of 5778 pairwise comparisons, respectively), and were the two most similar species of Pellonulini (1st out of 28 pairwise comparisons). When considering non-coding regions, the Tanganyika species pair was only the 22nd most similar, with *L. miodon* being more similar to most other pellonulines than to *S. tanganicae*. Among-group comparisons showed that *L. miodon* and *S. tanganicae* were almost equally differentiated from the remaining pellonulines when considering PCGs only (distance ± SE for *S. tanganicae* = 0.208 ± 0.006, *L. miodon* = 0.210 ± 0.006), but that *S. tanganicae* was more than twice as differentiated when considering non-coding regions only (distance ± SE for *S. tanganicae* = 0.221 ± 0.011, *L. miodon* = 0.106 ± 0.005).

Phylogenetic analysis

Maximum likelihood (ML) and Bayesian analysis (Figs. 3, 4) placed *S. tanganicae* and *L. miodon* together with the other members of Pellonulini with high statistical support. Within Pellonulini, several genera appeared non-monophyletic. The position of *Microthrissa royauxi* was unresolved. The Lake Tanganyika sardines formed a well-supported clade nested within *Potamothrissa*, while *M. congica* clustered with members of *Pellonula* and *Odaxothrissa.* In the latter clade, *P. vorax* was placed closer to *O. vittata* than to *P. leonensis* or to *O. losera.*

Outside of Pellonulini, all subfamilies of Clupeiformes, except Clupeinae, and most of their genera were retrieved with high support. Several deeper node placements in both trees, including some of the traditional clupeiform families with low taxonomic coverage, such as Pristigasteridae, Dussumieridae and Clupeidae II, had low support. Overall, ML and Bayesian analyses were in agreement, with the exception of those deeper, poorly supported nodes. We also found some genera split up or ambiguously placed. For example, there was a closer relationship between *Lycothrissa crocodilus* and *Setipinna melanochir* than the latter with other species of *Setipinna*. *Thryssa baleama* also did not cluster with other representatives of its own genus. *Ilisha elongata* clustered with *Pellona ditchela* and *Opisthopterus tardoore*, while *I. africana* and *I. sirishai* branched off earlier. Only two of the three species of *Sprattus* clustered together. The third, *S. sprattus*, was sister to the clade consisting of the two species of *Clupea*.

Dating of divergence time

Bayesian analysis estimated the divergence time of the Lake Tanganyika sardines at 4.07 MYA [95% CI: 1.66, 7.06] and the divergence between the most recent common ancestor (MRCA) of the Tanganyika sardines and their closest living relative in the tree, *P. obtusirostris*, at 11.88 MYA [95% CI: 6.09–17.2]. The split between Pellonulini and the other clupeids and thus the timing of a large marine incursion into north-western Africa, was estimated at 45.56 MYA [95% CI: 32.70-58.91] (Table 1, Fig. 5).

Table 1
Comparison of key divergence times, taxa and markers in the pellonuline phylogeny between studies.

	Divergence time estimate (MYA)				
	This study	Egan et al. 2018	Bloom & Lovejoy 2014	Wilson et al. 2008	Lavoué et al. 2013
Markers	**PCGs**	**CYT-B**, 16S, rag1, rag2, slc, zic1	**16S, CYT-B**, rag1, rag2	**16S, 12S, CYT-B**	PCGs, tRNAs, rRNAs
Number of taxa (excl. outgroup)	108	190	153	49	82
Number of sites (bp)	18106	7135	5211	1049–1811	10733
Node					
Limnothrissa miodon - Stolothrissa tanganicae	4.07 [1.66–7.06]	3.91 [1.19–6.64]	6.61 [2.20–11.01]	7.6 [2.1–15.9]	-
LT sardines – *Potamothrissa obtusirostris*	11.88 [6.09–17.27]	10.04 [5.62–14.47]	23.35 [16.37–30.33]	-	-
LT sardines – other pellonulines	-	-	-	27 [25.0–53.3]	-
Incursion 1: pellonulines - other clupeids	45.56 [32.70–58.91][1]	34.30 [25.56–43.03][1]	47.58 [35.68–59.47][1]	37 [25.0–53.3][2]	46.05 [33.38–58.71][1]
Incursion 2: *Gilchristella - Sauvagella*	-	25.00 [13.39–36.61]	33.92 [18.94–48.90]	20 [7.5–34.4]	-
Ehiravini - Pellonulini	70.49 [57.95–83.59]	70.13 [59.74–83.44]	98.24 [85.02–111.46]	48 [34.0–66.2]	89.02 [80.97–97.08]

Numbers between square brackets indicate 95% credibility intervals. Divergence times from our study were estimated in BEAST, those from other studies were directly reported or extracted from time-calibrated trees using WebPlotDigitizer. Markers indicated in bold were available for both *Stolothrissa tanganicae* and *Limnothrissa miodon*. PCGs = all mitochondrial protein coding genes, CYT-B = cytochrome B, 16S = 16S rRNA, 12S = 12S rRNA. [1] Split Pellonulini – Ethmalosa fimbriata. [2] Split Pellonulini – other clupeids (E. fimbriata not included in the study)

We verified the robustness of our secondary calibration based on node ages from Hughes et al. (36). Our estimates of fossil node ages corresponded to the ages from the literature. **F1: †***Cynoclupea nelsoni* as MRCA of Clupeoidei was estimated at 131.87 [95% CI: 106.25, 157.90], within the fossil-based age of minimum 125 MYA (soft 95% maximum 145 MYA). The second fossil node, **F2: †***Eoengraulis fasoloi* as the MRCA of Engraulidae, was estimated at 66.25 MYA [95% CI: 52.5, 80.0], corresponding to the age of the fossil with a minimum age of 50 MYA and a soft 95% maximum age of 86.3 MYA. **F3**: *Dorosoma petenense* was estimated as relatively old at 19.37 MYA [95% CI: 5.62, 33.12], but again within the boundaries of the true fossil age of minimum 2.5 MYA (soft 95% maximum 86.3 MYA).

Discussion

We used NGS to sequence and assemble the complete mitochondrial genomes of the Tanganyika sardines, *S. tanganicae* and *L. miodon*, and built a phylogeny of Clupeiformes using full mitochondrial sequences with a focus on the West and Central-African tribe Pellonulini. Based on these complete mitogenomes, we estimated the divergence time of the Tanganyika sardines to investigate the timing of their speciation in relation to the geology of Lake Tanganyika.

Conserved gene order in Clupeiformes

Generally, mitochondrial gene arrangements have remained stable for long evolutionary times, but rearrangements do occur in many lineages of both invertebrates and vertebrates. Small rearrangements of neighbouring genes, for example clusters of tRNA-genes, and non-coding regions are especially common (37–39). Several lineages of Actinopterygii are characterized by such rearrangements, but Clupeiformes is not one of them (35). Our gene order analysis confirmed the conserved arrangement of mitochondrial genes in this order, aside from one transposition of two tRNA genes (tRNA-Pro and tRNA-Thr) in *Ilisha elongata.*

Inconsistent tree topologies within and outside Pellonulini

Both of our phylogenies (ML and Bayesian) support a single common ancestor for all included pellonuline species and recover *Ethmalosa fimbriata* as their sister species with high statistical support, consistent with Lavoué et al. (21). This is in contrast with the results of Egan et al. (20) and Bloom & Lovejoy (19), neither of whom found good support for this sister-species relationship. None of the genera *Microthrissa*, *Odaxothrissa*, or *Pellonula* appeared monophyletic in our study, and the Tanganyika sardines rendered *Potamothrissa* paraphyletic. The position of *Microthrissa* differs in almost every study that has addressed it. According to Egan et al. (20), *M. royauxi* was more closely related to the Tanganyika sardines and *Potamothrissa*, while *M. congica* clustered with *Pellonula* and *Odaxothrissa.* In the study of Wilson et al. (17), two specimens of *M. royauxi* did not even cluster together. Bloom & Lovejoy (19), on the other hand, found both *Microthrissa* species more closely related to members of *Pellonula* and *Odaxothrissa* than to the clade including the Tanganyika sardines and *Potamothrissa.* In accordance, our analysis failed to recover the exact position of *M. royauxi* with high statistical support, but it did confirm that *M. congica* is more closely related to members of *Pellonula* and *Odaxothrissa* than to *M. royauxi.* The Lake Tanganyika sardines formed a well-supported clade nested within *Potamothrissa*, while *M. congica* clustered with members of *Pellonula* and *Odaxothrissa.* In the same clade, *P. vorax* and *O. vittata* were more closely related to each other than to *P. leonensis* or *O. losera.* Contrarily, our study is the first to place *P. obtusirostris* and *P. acutirostris* in the same clade. With the improved resolution resulting from our whole mitogenome approach, our study confirms *E. fimbriata* as sister species of Pellonulini and for the first time opposes monophyly of nearly all pellonuline genera.

The source of these inconsistent topologies is unclear, but could be related to smaller, more variable and partly incomplete gene datasets in previous studies. Egan et al. (20) included only the gene coding for CYT-B for most members of Pellonulini, including the LT sardines, and three nuclear genes and/or the 16S rRNA gene for others. Bloom & Lovejoy (19) did not include *O. losera* and *P. acutirostris* and used CYT-B and 16S rRNA genes for most, and two nuclear genes for two species, while Wilson et al. (17) used only mitochondrial genes (CYT-B, 16S rRNA, 12S rRNA). None of the previous studies included more than around 5 kbp of alignment data, except Lavoué et al. (21), who included all PCG, rRNA and tRNA sequences which amounted to slightly over 10kbp. Taxonomic coverage was also highly variable, ranging from 49 to 190 clupeoid species, the lowest number belonging to Wilson et al. (17), which also had the largest credibility interval but had the highest taxonomic coverage of Pellonulini. The Tanganyika sardines and other pellonulines were also missing from several of these studies. Although the inclusion of taxa with incomplete datasets can help to resolve phylogenies (40), there is a trade-off with increased risk of phylogenetic artefacts, and difficulty detecting multiple substitutions (41,42). In contrast, our study had the first nearly complete dataset for all taxa, thanks to the readily available mitochondrial genomes from many pellonuline and other clupeid species.

Morphological diversity is relatively low in representatives of Pellonulini compared to for example cichlids (15,43). In the FAO species catalogue of clupeoid fishes, several ambiguous identifications and uncertain species descriptions are mentioned. For instance, the distinction between *O. losera* and *O. vittata* is based solely on the number of gill rakers, which also varies with the age of the specimen, a common occurrence among clupeid fishes (43). In *P. leonensis*, there is also evidence for undescribed subspecies exhibiting characteristics of both *P. leonensis* and *P. vorax*, or even specimens belonging to *Cynothrissa.* It is thus not inconceivable that misidentifications have confounded past taxonomic studies, and that a taxonomic revision of these genera may be needed.

Outside of Pellonulini, we recovered most of the traditional families and subfamilies of Clupeiformes with high statistical support. The positions of the families with lower taxonomic coverage, including Pristigasteridae, Chirocentridae, Dussumieridae, remained unresolved, Clupeidae was not monophyletic and several species were also placed away from their congeners, for example in the genera *Setipinna*, *Thryssa, Ilisha* and *Sprattus.* These latter two findings are consistent with previous studies of Clupeiformes with taxonomic coverage comparable to ours (19,20,44), underlining the need for revision of these taxa.

Inconsistent divergence time estimates in Clupeiformes

Bayesian analysis estimated the divergence time of the Lake Tanganyika sardines at around 4.07 MYA [95% CI: 1.66, 7.06]. This estimate is younger than the previous estimate by Wilson et al. (17) at 7.6 MYA, but within its credibility interval [95% CI: 2.1, 15.9]. Conversely, their estimate fell just outside our credibility interval. Our estimate is also younger compared to the one by Bloom & Lovejoy (19) at 6.61 MYA [95% CI: 2.20, 11.01], but in accordance with Egan et al. (20) at 3.91 MYA [95% CI: 1.19, 6.64]. Our credibility intervals were smaller than those of Wilson et al. (17), but comparable to Bloom & Lovejoy (19) and Egan et al. (20). Other, deeper nodes of interest also differed between the studies. Overall, our estimates most closely agreed with those of Egan et al. (20), except for the divergence time of the pellonulines from other clupeids, which was around 10 MYA older in our study. Bloom & Lovejoy (19) found consistently older estimates, while Wilson et al. (17) estimated the more recent nodes as older, and the deeper nodes as younger than the other three studies.

The differences in estimated divergence times between the studies can be partially attributed to the different estimation procedures. Specifically, methodological choices for Bayesian dating of nodes can strongly influence the accuracy and precision of the divergence time estimates, for example the choice of priors to account for uncertainty surrounding the age of a fossil, and the choice of clock model (34,45). Almost all the studies we compared here dated divergence using a fossil-calibrated uncorrelated (relaxed) clock model implemented in BEAST, accounting for substitution rate heterogeneity among branches. Six to eight fossil calibrations were specified as exponential priors with soft maximum ages. Only Wilson et al. (17) used an autocorrelated clock approach with seven fossil calibrations specified as uniform priors. In accordance with the three more recent studies on the divergence times of Clupeiformes (19–21), but in contrast with Wilson et al. (17), we chose a relaxed clock model, in accordance with the varying speeds of diversification in different clupeid lineages (44). However, we decided to use pre-calibrated time scaling points ("secondary calibration") over direct fossil calibration ("primary calibration"). Ideally, time-calibration of the diversification of the pellonulines would be based on fossils within this clade. Unfortunately, there are no known pellonuline fossils, and the fossil record of fish of Central Africa in general is sparse (16). Secondary calibrations can result in younger and falsely narrowed estimates of node ages (42,46). By basing secondary calibrations on 95% credibility intervals of primary estimates, 5% of the primary uncertainty (the values falling outside of the credibility interval), is lost (42). Nevertheless, secondary calibrations were necessary in our case to achieve convergence. In addition, the low number of fossil calibration points that was applicable to our dataset would likely have resulted in less accurate estimates than using secondary estimates from carefully calibrated phylogenies (47). Our use of normally distributed priors (a common approach for secondary calibrations) rather than lognormal or exponential distributions could have shifted node age estimates in the opposite direction to produce older estimates (46). Considering these effects, we are confident that our methodological choices have produced node age estimates with the best possible accuracy, but some caution is warranted when interpreting the width of our credibility intervals. The robustness of our approach is supported by correspondence of our estimates to the applied fossil ages (potential primary calibrations).

Utility of whole mitochondrial genomes for phylogenomic analysis and divergence time dating

Mitochondrial protein coding genes vary in their ability to recover known phylogenetic topologies. The sequences of ND4, ND5, COI and CYTB genes are generally useful for phylogenetic questions, while fast evolving genes such as ND4L and ATP8 are regarded as poor phylogenetic performers, although this differs per study and taxon (25,26). Indeed, we found relatively high nucleotide diversity in some genes or regions compared to others, including parts of the genes coding for the ATP6, ND1, ND2, ND3 and parts of ND5. However, whole mitochondrial genomes can recover accurate phylogenies with high resolution, despite containing “poor” phylogenetic performers (27,48). A smaller subset of “good” mitochondrial genes may be able to recover the same topology as the entire mitochondrial genome, but this is highly taxon-specific (25–27,48). Thus, utilizing more markers that provide complementary information is preferable if previous taxon-specific information on the utility of single markers is not available (27).

Nuclear and mitochondrial DNA can contrast or complement each other, both in terms of tree topology and branch lengths (7,8,44). Mitochondrial data has been extensively used due to its uniparental inheritance as a single linkage group, little recombination and fast evolutionary rate, but has been criticized due to frequent violation of the selective neutrality assumption and complications related to introgression and incomplete lineage sorting between species (24). Our estimated divergence time of 4.07 MYA might indicate the lower boundary of divergence time if, after initial isolation, secondary contact with gene flow has occurred between the two diverging sardine species. Due to haploidy and uniparental inheritance, the effective population size of mitochondrial DNA (mtDNA) is fourfold smaller compared to nuclear DNA (nDNA) (24), implying that lineage sorting will happen more rapidly in mtDNA compared to in nDNA (49–51). Due to a lack of recombination, mtDNA introgresses as one single block. Therefore it is possible that through introgression, followed by complete lineage sorting, the mtDNA of one of the species replaces the mtDNA of the other species without evidence of introgression (mitochondrial capture) (51–53). In this case, our estimate would indicate the divergence time after mitochondrial homogenization, instead of the original divergence time. There are several examples of ongoing hybridization between clupeid species (54–57). With their similar habitat, nursery areas, and modes of reproduction, the Tanganyika sardines may well have a history of introgression. A recent study using RAD-tag sequencing suggested that they have completely different sex chromosome organization (58), making this prospect less likely, but not impossible (59).

At the start of the Eocene (around 50 MYA), global sea levels were more than 100 m higher than today, and steadily decreased over the next 20 million years, with smaller maxima in between (16,60,61). These fluctuations may have allowed frequent isolations and reconnections in the Congo basin, favouring hybridization between other newly formed pellonuline species as well. To completely resolve the species tree of Pellonulini, phylogenomic analyses using nuclear genomic markers and multiple individuals per species are needed (but see Bloom & Egan (44), who found similar divergence time estimates with mtDNA and nDNA datasets).

Recent divergence time suggests intralacustrine speciation of the Tanganyika sardines

Present-day distributions of several Afrotropical freshwater fish lineages show striking overlap, including members of Pellonulini, Kneriidae and Phractolaemidae, providing evidence for a single marine-freshwater transition across West- and Central Africa around 50 MYA during a period of high sea levels (16). Despite the high sea levels, Lake Tanganyika was likely never in direct contact with the ocean and has not experienced much higher water levels than at present (3,6). Furthermore, due to uplift of the borders of the Congo basin from the Cenozoic onwards, the possibility of an additional marine incursion close to the lake is faint (18). It is therefore more likely the Lake Tanganyika sardines evolved from riverine clupeids. Indeed, the presence of a large body of water covering a large area of the Congo basin ("paleo-lake Congo") until the Pliocene or early Pleistocene (2–12 MYA, Beadle, 1974; Peters and O'Brien, 2001), may have increased the connectivity between the Congo tributaries and its surrounding lakes, and may have facilitated the entry of riverine species into the predecessor of Lake Tanganyika at this time.

Our improved divergence time estimates of the Tanganyika sardines (4.07 MYA) and their MRCA from other pellonulines (11.88 MYA) help us to better understand their origin and colonisation time in connection to the geological history of the lake. Our estimates are compatible with (1) the entrance of the MRCA of the Tanganyika sardines into the newly formed Tanganyika basin (around 12 MYA) via the tributaries of the proto-Malagarasi-Congo river; and (2) intralacustrine speciation at the onset of deep- and clearwater conditions after the subbasins fused (5–6 MYA). However, based on the 95% credibility intervals of our estimates, we cannot exclude the possibility that the MRCA of the Tanganyika sardines diverged from *P. obtusirostris* outside of the proto-lake and entered it sometime between the time of its formation and the fusion of its subbasins.

Which environmental conditions triggered the divergence between *S. tanganicae* and *L. miodon* remains uncertain. Sexual selection, such as in cichlids (64), is unlikely to have played a large role due to the mode of reproduction of the clupeids. Ecological differences can be powerful drivers of speciation, even in (partial) sympatry (65,66). The newly fused basin, adding ecological heterogeneity to the ancestral sardine's environment, may have favoured dietary specialization through divergent selection on polymorphic trophic traits. Niche separation and divergence can then prompt genetic reproductive isolation if reinforced by spatial or temporal separation of spawning or lower hybrid fitness (65–67). Indeed, contemporary populations of *L. miodon* seem to spawn all year round and mostly in the littoral, while populations of *S. tanganicae* exhibit clear peak spawning times in the pelagic (68). This suggests that at some point during their divergence, spawning became more common in their respective preferred habitats. An alternative explanation is that their speciation was triggered by periods of allopatry (65,67). Given our credibility intervals and the frequent water-level fluctuations potentially separating and reconnecting the southern and central subbasins of Lake Tanganyika several times, it is likely that ancestral sardine populations frequently occurred in partial or complete isolation.

A *Limnothrissa*-like ancestor of the Tanganyika sardines?

According to our ML and Bayesian phylogenies, the closest living relative of the Tanganyika sardines is *P. obtusirostris*, a riverine herring feeding primarily on insects that occurs in the northern and eastern stretches and tributaries of the Congo river system, all the way down to the Lukuga river, which was connected to the Malagarasi river east of Tanganyika around the time of the lake's formation. Ecologically, *L. miodon*, with its generalist diet including insects and small fishes, is more similar to *P. obtusirostris* than *S. tanganicae*, which is a strict planktivore. In addition, individuals of *L. miodon* and species of *Potamothrissa* share a morphological feature that is otherwise rare in clupeid fishes: a row of saw-like teeth at the side of the lower jaw (43). We suggest that the ancestral Tanganyika sardine shared more ecological traits with *L. miodon* than with *S. tanganicae.* This is also reflected in the more shorebound and generalist lifestyle of *L. miodon*, and its ability to invade the Cahora Bassa reservoir though dispersal via the riverine environment of the Zambezi (69), suggesting a relatively high ecological flexibility compared to *S. tanganicae* (70), and thus a higher ability to colonize a new environment. Nevertheless, the presence of established contemporary populations of *S. tanganicae* in one of the Congo's tributaries, the Lukuga, attest its ability to inhabit, or at least cross, non-pelagic environments, provided the water composition is sufficiently similar (71). We also found larger genetic differentiation of *S. tanganicae* than *L. miodon* from the remaining pellonulines in non-coding regions. This could further support our hypothesis of a higher relatedness between *L. miodon* and the ancestral sardine, but may also indicate different demographic histories (72). Kmentová et al. (73) found signatures of recent population expansion in both *L. miodon* and *S. tanganicae*, but these were more pronounced in the latter. The population expansion in *S. tanganicae* might be linked to the fusion of subbasins, or any other major lake-level fluctuation that increased the amount of pelagic habitat. Similarly, species of the pelagic cichlid tribe Bathybatini showed recent demographic expansions, probably also linked to lake-level fluctuations (74).

Conclusion

Using NGS data, we assembled and annotated the full mitochondrial genomes of the Tanganyika sardines *S. tanganicae* and *L. miodon.* Putting them into phylogenetic context with full mitochondrial genomes of 109 other clupeid species, we estimate their divergence time at 4 MYA, and divergence from their riverine ancestor at 12 MYA. This relatively young estimate implies that the MRCA of the Tanganyika sardines entered Lake Tanganyika shortly after its formation during a period of high connectivity of the Congo basin's water bodies. We suggest that the speciation event is likely to have been brought on by the fusion of Lake Tanganyika's subbasins and the subsequent clear water conditions.

The mitochondrial genomes of *S. tanganicae* and *L. miodon* are valuable resources for future studies of the evolutionary history of these species at the population level, for example as a reference for barcoding, studies of their mitochondrial diversity and evolutionary history, as well as macroevolutionary study of relationships within Pellonulini and Clupeiformes. Future work should focus on the divergence time of different regions of the Tanganyika sardines' genomes and compare them to a dataset of nuclear genes or genome-wide data. This in combination with formal tests for hybridization could help to gauge the role of introgression in the timing and the scenario of speciation. Nuclear genomic sequences from several individuals of all members of Pellonulini would allow a more precise reconstruction of their colonization of West-Africa and clarify the ambiguous classifications in this group.

Methods

DNA extraction, library preparation and sequencing

One female individual of the two species was collected from liftnet fishing catches on the night of 15th of December 2018 off the shore of Uvira, Democratic Republic of Congo. Fish were dissected to extract liver tissue, which was directly frozen on dry ice and subsequently stored at -20°C until extraction. High molecular weight genomic DNA (gDNA) was extracted using a Blood and cell culture DNA Midi kit (Qiagen). Libraries were prepared for each species separately using Chromium Genome Library & Gel Bead Kit v.2 (10X Genomics, cat. 120258), Chromium Genome Chip Kit v.2 (10X Genomics, cat. 120257), Chromium i7 Multiplex Kit (10X Genomics, cat. 120262) and Chromium controller according to the manufacturer’s instructions with one modification (added shearing step before Illumina library prep). Briefly, gDNA diluted to 1.02 ng/ µl was combined with Master Mix, a library of Genome Gel Beads, and partitioning oil to create Gel Bead-in-Emulsions (GEMs) on a Chromium Genome Chip. The GEMs were isothermally amplified with primers containing an Illumina Read 1 sequencing primer, a unique 16 bp 10X barcode and a 6 bp random primer sequence. Barcoded DNA fragments were recovered for Illumina library construction. The amount and fragment size of post-GEM DNA was quantified using a Bioanalyzer 2100 with an Agilent High sensitivity DNA kit (Agilent, cat. 5067 – 4626). Prior to Illumina library construction, the GEM amplification product was sheared on an E220 Focused Ultrasonicator (Covaris, Woburn, MA) to approximately 350 bp (55 seconds at peak power = 175, duty factor = 10, and cycle/burst = 200). Then, the sheared GEMs were converted to a sequencing library following the 10X standard operating procedure. The library was quantified by qPCR with a Kapa Library Quant kit (Kapa Biosystems-Roche) and sequenced on a partial lane of the NovaSeq6000 sequencer (Illumina, San Diego, CA) with paired-end 150 bp reads.

Mitochondrial genome assembly

For mitogenome assembly, raw 10X Chromium reads were processed using the proc10xG package (75). Process_10xReads.py was run using default settings to remove GEM and individual sample barcodes. The resulting reads passed read assessment by FastQC v0.11.7 (76) without any quality problems, residual adapters or overrepresented sequences, the latter also commonly indicating adapter contamination. Mitogenomes of the two sardines were assembled from these barcode trimmed reads using MitoZ v.2.4-alpha (77) with default settings. Mitochondrial genes were annotated using the Mitofish annotator web service (78).

Taxonomic sampling and alignment

Taxonomic sampling for phylogenetic analysis included all members of Clupeiformes for which a complete mitochondrial genome is published (accessed 21st of October 2020, see Supplementary Table 1, Additional file 1). The sequences of *Odaxothrissa vittata* (NC_009590.1) and *Etrumeus teres* (NC_009583.1) were identical to *Pellonula vorax* and *E. micropus*, respectively, and were omitted from analysis. The outgroup was selected based on Lavoué et al. (2013) (21) and includes the denticle herring

(*Denticeps clupeoides*), two alepocephaliforms, four ostariophysians and two euteleosts (Supplementary Table 1, Additional file 1). We extracted mitogenomes and their annotations from NCBI using a combination of efetch (79) and custom bash, perl and python scripts. We manually verified the new annotations of *S. tanganicae* and *L. miodon* by comparing the nucleotide and amino acid sequences, translated using vertebrate mitochondrial code, to the already published pellonuline mitogenomes, and checking for the presence of start- and stop-codons at the appropriate positions in MEGA-X v.10.0.5 (80).

We separately aligned each PCG using a codon-based MAFFT algorithm in the TranslatorX server (81). We selected options for less stringent selection which allowed smaller final blocks with gap positions and less strict flanking positions. D-loop sequences were aligned using MAFFT v.7.470 with default parameters, and sequences of the rRNA coding regions using MAFFT with the -qinsi option (82). Incomplete or missing regions were coded as missing data (N). We performed all alignments with and without alignment cleaning by Gblocks, further referred to as ‘complete’ and ‘trimmed’ alignments. We used AMAS v.0.98 (83) to separately concatenate the trimmed and complete alignments, producing one trimmed and one complete dataset. The phylogenetic content of these datasets was compared using likelihood mapping (84) implemented in TREE-PUZZLE v.5.3.rc16 (85). We determined the optimal model for each dataset using jmodeltest2 (86) and specified these as input models for TREE-PUZZLE. Since there was no difference in phylogenetic content (86.6% fully resolved quartets, 2.3% partly resolved, 11.1% unresolved), we performed all subsequent analyses on the complete dataset of 18106 bp.

Genetic diversity, divergence, and gene order

We calculated nucleotide diversity (π) of the final alignment using a sliding window analysis implemented in DNAsp v.6 (87) with a window size of 300 bp and steps of 15 bp. We used MEGA-11 (88) to quantify divergence between species and clades. We calculated pairwise genetic distances between all species, and mean between-group genetic distances between *S. tanganicae, L. miodon* and the remaining members of Pellonulini and Clupeiformes (in each of these comparisons excluding the other Tanganyika clupeid). The distances were calculated separately for PCGs (vertebrate mitochondrial code) and non-coding regions using a Tamura-Nei model including transitions and transversions, gamma-distributed rate variation among sites and heterogenous rate patterns among lineages. Gaps and missing data were deleted in a pairwise manner. The gamma parameter was estimated separately for the PCG and non-coding dataset using jmodeltest2. To estimate the relative similarity of *S. tanganicae* and *L. miodon* compared to similarities among other clupeids, we ranked all pairwise genetic distances of 1) Clupeiformes and 2) Pellonulini and calculated in which percentile the Tanganyika sardines fell using R v.4.0.4 (89). Finally, we compared the gene order of all species included in our study using the CREx web application (90). Two species, *Alosa fallax* and *Hilsa kelee*, were missing several markers (genes), and were thus excluded from this analysis.

Phylogenomic tree building

We ran IQ-TREE v.1.6.12 (91) twice on the complete dataset. The first run determined the best partition scheme (option -m MF + MERGE), allowing different models of molecular evolution in different genes/regions and, for PCGs, at the different codon positions (29). The second run first implemented ModelFinder (92) to find the optimal model of evolution for each partition found by the previous round (options -spp and -m MFP), then constructed a ML tree, and finally assessed nodal support by 10,000 ultrafast bootstraps (generating support value UFBoot%) and 1,000 Shimodaira-Hasegawa-like approximate likelihood ratio test replicates (generating support value SH-aLRT%). A clade can be considered well supported if UFBoot% ≥ 95 (corresponding to a ~ 95% chance that the clade is true) and SH-aLRT% ≥ 80 (93,94).

Using the IQ-TREE-derived partition and models, we constructed a Bayesian phylogeny in MrBayes v.3.2.7a (95), allowing estimation of the model parameters for each partition separately (unlinked character state frequencies, substitution rates of the General Time Reversible model, proportion of invariable sites and gamma shape parameter). Two independent runs with 4 Markov Chain Monte Carlo chains ran for 60 million generations, sampling every 500 generations and discarding the first 25% as burn in. The remaining samples were used to calculate Bayesian posterior probabilities (BPP) for each node in order to assess nodal support. A clade is considered well supported if BBP ≥ 0.9. The models converged, as indicated by the average standard deviation of split frequencies approaching zero, the absence of a trend in log likelihood of the runs, an Effective Sample Size (ESS) > 200, and the Potential Scale Reduction Factor approaching 1 (96).

Dating of divergence time

We estimated branch lengths and divergence times between *S. tanganicae* and *L. miodon* and five other nodes of interest (Table 1) using Bayesian relaxed molecular clock analysis implemented in BEAST v.2.6.3 (97) with the tree topology from MrBayes as a starting tree. We conducted two independent BEAST runs of 50 million generations. Convergence was ensured by checking if ESS was higher than 200 for all parameters. Results of the BEAST runs were analyzed using Tracer v.1.7.1 (98). Trees were summarized and annotated using the TreeAnnotator module in BEAST, and the final time-calibrated tree was visualized using FigTree v.1.4.4 (99).

For the time calibration of our tree, we first attempted a “primary calibration”, which relies solely on fossils, using the complete dataset. However, when this model failed to converge, even after hundreds of millions of generations, we modified our analysis in two ways. First, we sequentially omitted the third codon positions, the D-loop sequence, and regions coding for rRNA, leaving a dataset with only first and second codon positions of all PCGs. We also merged the separate partitions found by IQ-TREE into only two partitions for the first and second codon positions, respectively. Second, we chose to apply “secondary calibration”, which uses estimates from an already existing phylogeny.

For the primary calibration, we chose a subset of fossils outlined by Bloom & Lovejoy, Egan et al., Lavoué et al. and Wilson et al. (17,19,20,100), based on applicability to our dataset. The calibration points were implemented as exponential prior distributions for the node ages in BEAST. **F1:** †*Cynoclupea nelsoni* (101) as the MRCA of all members of Clupeiformes (including *D. clupeoides*) with a minimum age of 125 MYA and a soft 95% maximum age of 145 MYA due to the absence of clupeoid fossils from the Jurassic and earlier. **F2: †***Eoengraulis fasoloi* (102) as the MRCA of the members of Engraulidae with a minimum age of 50 MYA and a soft 95% maximum age of 86.3 MYA. **F3**: Earliest known *Dorosoma petenense* fossil (103) as the MRCA of *D. petenense* and *D. cepedianum* with a minimum age of 2.5 MYA and a soft 95% maximum age of 86.3 MYA because most crown group clupeoids are younger than the limit Coniacian/Santonian.

For the secondary calibration, we used three calibration points from a recent phylogeny of the teleosts using more than 30 fossils (36). We included the MRCA of members of Clupeiformes and our outgroup (including *Danio rerio, Cyprinus carpio* and *Chanos chanos*) at 194 MYA (**C1**), MRCA of members of Clupeinae and Alosinae at 73 MYA (**C2**), and MRCA of Engraulidae at 61 MYA (**C3**). The calibration points were implemented as normal prior distributions for the node ages in BEAST.

Secondary calibration inevitably incorporates geological and fossil uncertainty along with uncertainties associated with the primary dataset. They tend to push node age estimates into the more recent direction and falsely narrow the credibility intervals, especially when using a single old secondary calibration (42,46), but see Powell et al. (47). In our case, we presume the associated problems to be minimal for four reasons: 1) We used secondary calibrations for both old and younger nodes, which should diminish the tendency to estimate other nodes as younger (46). 2) In Bayesian analysis implemented in BEAST, lognormally or exponentially distributed prior distributions are most commonly used for primary calibrations. These produce younger node age estimates and narrower credibility intervals than uniform (and presumably normally distributed) priors (46), which are more commonly used for secondary calibrations. 3) We validated the robustness of our secondary calibration by comparing the results of the BEAST dating to our chosen primary fossil calibration points. 4) The true uncertainty associated with secondary calibrations from Hughes et al. (36) based on more than 30 well-characterized fossils is likely smaller than that produced by using only three or four primary fossils with lognormal distributions with large variance.

Finally, we compared our divergence time estimates for six nodes with published estimates (17,19–21) (Table 1). Estimates and 95% CI that were not directly reported in these publications were extracted from time-calibrated trees using WebPlotDigitizer v.4.5 (104).

Abbreviations

MYA
Million Years Ago
COI
cytochrome c oxidase subunit I
CYTB
cytochrome B
rRNA
ribosomal RNA

ND

NADH-ubiquinone oxidoreductase

ATP6

ATP synthase membrane subunit 6

mtDNA

Mitochondrial DNA

nDNA

Nuclear DNA

gDNA

Genomic DNA

GEMs

Gel Bead-in-Emulsions

tRNA

Transfer RNA

NGS

Next generation sequencing

PCGs

Protein-coding genes

ML

Maximum likelihood

MRCA

Most recent common ancestor

CI

credibility interval

BPP

Bayesian posterior probabilities

ESS

Effective Sample Size

CRH-U

Centre de Recherche en Hydrobiologie – Uvira

Declarations

Ethics approval and consent to participate

Collection of specimens used in the study complied with institutional, national, and international guidelines. Fieldwork in the Democratic Republic of Congo was carried out with the approval of the Centre de Recherche en Hydrobiologie – Uvira (CRH-U), which falls under the Congolese Ministry for science and technology ("Ministère National de la Recherche Scientifique et Technologie") under mission statement 002/MINRST/CRH-U/2018. Samples were exported with an export permit from the CRH-U. Samples were imported the USA under import permit 70881B by the Department of the Interior U.S. Fish and Wildlife Service, Office of Law Enforcement to Dovetail Genomics. No animals were killed for this study, dead specimens were obtained from fishermen on Lake Tanganyika.

Consent for publication

Not applicable.

Availability of data and materials

The datasets supporting the conclusions of this article are available in the NCBI Sequence Read Archive [http://www.ncbi.nlm.nih.gov/bioproject/860551] and GenBank [Accessions: OP022425, OP021863] repositories. Custom scripts are available on Github: https://github.com/lmilec/mitoprep

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Dovetail genomics for sequencing of the NGS data, and Armando J. Cruz-Laufer and Michael Matschiner for advice on analysis and visualization.

Funding

LJMM is supported by her Nord University PhD grant 224000-131. LJMM and JAMR are further supported by JAMR's start-up grant at Nord University, covering sequencing costs. The Special Research Fund of Hasselt University finances LJMM (BOF19DOC31) and MPMV (BOF20TT06). MPMV is further supported by Czech Science Foundation standard project GA19-13573S. ELRDK was funded through VLIR-UOS (VLADOC scholarship NDOC2016PR006 and South Initiative project CD2018SIN218A101) and BOF Seal of Excellence of University of Antwerp.

Author contributions

LJMM, MPMV and JAMR conceptualized the study. LJMM collected and analysed the sequencing data and wrote the draft with input and interpretation by MPMV and JAMR. FMB, ELRDK, VLK, PMM and NM coordinated and executed the sampling of the Tanganyika sardines. All authors read and revised the draft and approved the final manuscript.

References

1. Salzburger W, Van Bocxlaer B, Cohen AS. Ecology and Evolution of the African Great Lakes and Their Faunas. Annu Rev Ecol Evol Syst. 2014;45(1):519–45.
2. Cohen AS, Soreghan MJ, Scholz CA. Estimating the age of formation of lakes: an example from Lake Tanganyika, East African Rift system. Geology. 1993;21(6):511–4.
3. Tiercelin J, Mondeguer A. The geology of the Tanganyika Trough. In: Coulter GW, editor. Lake Tanganyika and its life. Oxford University Press; 1991. p. 7–48.
4. Tiercelin J-J, Lezzar K-E. A 300 Million Years History of Rift Lakes in Central and East Africa: An Updated Broad Review. In: Odada EO, Olago DO, editors. The East African Great Lakes: Limnology, Palaeolimnology and Biodiversity Advances in Global Change Research, vol 12. Springer, Dordrecht; 2002. p. 3–60.
5. Cohen AS, Stone JR, Beuning KRM, Park LE, Reinthal PN, Dettman D, et al. Ecological consequences of early Late Pleistocene megadroughts in tropical Africa. Proc Natl Acad Sci U S A. 2007;104(42):16422–7.
6. Cohen AS, Lezzar KE, Tiercelin A JJ, Soreghan M. New palaeogeographic and lake-level reconstructions of Lake Tanganyika: Implications for tectonic, climatic and biological evolution in a rift lake. Basin Res. 1997;9(2):107–32.
7. Salzburger W, Meyer A, Baric S, Verheyen E, Sturmbauer C. Phylogeny of the Lake Tanganyika cichlid species flock and its relationship to the Central and East African haplochromine cichlid fish faunas. Syst Biol. 2002;51(1):113–35.
8. Sturmbauer C, Salzburger W, Duftner N, Schelly R, Koblmüller S. Evolutionary history of the Lake Tanganyika cichlid tribe Lamprologini (Teleostei: Perciformes) derived from mitochondrial and nuclear DNA data. Mol Phylogenet Evol. Elsevier Inc.; 2010;57(1):266–84.
9. Meyer BS, Matschiner M, Salzburger W. Disentangling Incomplete Lineage Sorting and Introgression to Refine Species-Tree Estimates for Lake Tanganyika Cichlid Fishes. Syst Biol. 2017;66(4):531–50.
10. Klett V, Meyer A. What, if anything, is a Tilapia? - Mitochondrial ND2 phylogeny of tilapiines and the evolution of parental care systems in the African cichlid fishes. Mol Biol Evol. 2002;19(6):865–83.
11. Koch M, Koblmüller S, Sefc KM, Duftner N, Katongo C, Sturmbauer C. Evolutionary history of the endemic Lake Tanganyika cichlid fish Tylochromis polylepis: A recent intruder to a mature adaptive radiation. J Zool Syst Evol Res. 2007;45(1):64–71.
12. Ronco F, Matschiner M, Böhne A, Boila A, Büscher HH, El Taher A, et al. Drivers and dynamics of a massive adaptive radiation in cichlid fishes. Nature. 2021;589(7840):76–81.
13. Meyer BS, Matschiner M, Salzburger W. A tribal level phylogeny of Lake Tanganyika cichlid fishes based on a genomic multi-marker approach. Mol Phylogenet Evol. Elsevier Inc.; 2015;83:56–71.
14. Irisarri I, Singh P, Koblmüller S, Torres-Dowdall J, Henning F, Franchini P, et al. Phylogenomics uncovers early hybridization and adaptive loci shaping the radiation of Lake Tanganyika cichlid fishes. Nat Commun. Springer US; 2018;9(1).

15. Gourène G, Teugels GG. Synopsis de la classification et phylogénie des Pellonulinae de l'Afrique Occidentale et Centrale (Teleostei; Clupeidae). J African Zool. 1994;108(1):77–91.

16. Lavoué S. Origins of Afrotropical freshwater fishes. Zool J Linn Soc. 2020;188(2):345–411.

17. Wilson AB, Teugels GG, Meyer A. Marine incursion: The freshwater herring of Lake Tanganyika are the product of a marine invasion into West Africa. PLoS One. 2008;3(4).

18. Giresse P. Mesozoic-Cenozoic history of the Congo Basin. J African Earth Sci. 2005;43(1–3):301–15.

19. Bloom DD, Lovejoy NR. The evolutionary origins of diadromy inferred from a time-calibrated phylogeny for Clupeiformes (herring and allies). Proc R Soc B Biol Sci. 2014;281(1778).

20. Egan JP, Bloom DD, Kuo CH, Hammer MP, Tongnunui P, Iglésias SP, et al. Phylogenetic analysis of trophic niche evolution reveals a latitudinal herbivory gradient in Clupeoidei (herrings, anchovies, and allies). Mol Phylogenet Evol. 2018;124(March):151–61.

21. Lavoué S, Miya M, Musikasinthorn P, Chen WJ, Nishida M. Mitogenomic Evidence for an Indo-West Pacific Origin of the Clupeoidei (Teleostei: Clupeiformes). PLoS One. 2013;8(2).

22. Coulter GW. Biomass, Production, and Potential Yield of the Lake Tanganyika Pelagic Fish Community. Trans Am Fish Soc. 1981;110(3):325–35.

23. Mölsä H, Reynolds JE, Coenen EJ, Lindqvist O V. Fisheries research towards resource management on Lake Tanganyika. Hydrobiologia. 1999;407:1–24.

24. Ballard JWO, Whitlock MC. The incomplete natural history of mitochondria. Mol Ecol. 2004;13(4):729–44.

25. Cummings MP, Otto SP, Wakeley J. Sampling properties of DNA sequence data in phylogenetic analysis. Mol Biol Evol. 1995;12(5):814–22.

26. Zardoya R, Meyer A. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol Biol Evol. 1996;13(7):933–42.

27. Duchêne S, Archer FI, Vilstrup J, Caballero S, Morin PA. Mitogenome phylogenetics: The impact of using single regions and partitioning schemes on topology, substitution rate and divergence time estimation. PLoS One. 2011;6(11).

28. Naylor GJP, Collins TM, Brown WM. Hydrophobicity and phylogeny. Nature. 1995;373(6515):566.

29. Chernomor O, Von Haeseler A, Minh BQ. Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices. Syst Biol. 2016;65(6):997–1008.

30. Lanfear R, Calcott B, Ho SYW, Guindon S. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 2012;29(6):1695–701.

31. Lv Y, Li Y, Ruan Z, Bian C, You X, Yang J, et al. The complete mitochondrial genome of *Glyptothorax macromaculatus* provides a well-resolved molecular phylogeny of the Chinese sisorid catfishes. Genes (Basel). 2018;9(6):1–13.

32. Yamanoue Y, Miya M, Matsuura K, Katoh M, Sakai H, Nishida M. A new perspective on phylogeny and evolution of tetraodontiform fishes (Pisces: Acanthopterygii) based on whole mitochondrial genome sequences: Basal ecological diversification? BMC Evol Biol. 2008;8(1):1–14.

33. Sullivan KAM, Platt RN, Bradley RD, Ray DA. Whole mitochondrial genomes provide increased resolution and indicate paraphyly in deer mice. BMC Zool. BMC Zoology; 2017;2(1):1–6.

34. Zhang P, Papenfuss TJ, Wake MH, Qu L, Wake DB. Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes. Mol Phylogenet Evol. Elsevier Inc.; 2008;49(2):586–97.

35. Satoh TP, Miya M, Mabuchi K, Nishida M. Structure and variation of the mitochondrial genome of fishes. BMC Genomics. BMC Genomics; 2016;17(1):1–20.

36. Hughes LC, Ortí G, Huang Y, Sun Y, Baldwin CC, Thompson AW, et al. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc Natl Acad Sci U S A. 2018;115(24):6249–54.

37. Boore JL. Animal mitochondrial genomes. Nucleic Acids Res. 1999;27(8):1767–80.

38. Dowton M, Castro LR, Austin AD. Mitochondrial gene rearrangements as phylogenetic characters in the invertebrates: The examination of genome “morphology.” Invertebr Syst. 2002;16(3):345–56.

39. Satoh TP, Sato Y, Masuyama N, Miya M, Nishida M. Transfer RNA gene arrangement and codon usage in vertebrate mitochondrial genomes: A new insight into gene order conservation. BMC Genomics. 2010;11(1).

40. Wiens JJ. Missing data and the design of phylogenetic analyses. J Biomed Inform. 2006;39(1 SPEC. ISS.):34–42.

41. Roure B, Baurain D, Philippe H. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol Biol Evol. 2013;30(1):197–214.

42. Schenk JJ. Consequences of secondary calibrations on divergence time estimates. PLoS One. 2016;11(1).

43. Whitehead PJP. Clupeoid fishes of the world (Suborder Clupeoidei): An annotated and illustrated catalogue of the herrings, sardines, pilchards, sprats, shads, anchovies and wolf-herrings. Rome, Italy: FAO; 1985.

44. Bloom DD, Egan JP. Systematics of clupeiformes and testing for ecological limits on species richness in a trans-marine/freshwater clade. Neotrop Ichthyol. 2018;16(3).

45. Battistuzzi FU, Filipski A, Hedges SB, Kumar S. Performance of relaxed-clock methods in estimating evolutionary divergence times and their credibility intervals. Mol Biol Evol. 2010;27(6):1289–300.

46. Sauquet H, Ho SYW, Gandolfo MA, Jordan GJ, Wilf P, Cantrill DJ, et al. Testing the impact of calibration on molecular divergence times using a fossil-rich group: The case of *Nothofagus* (Fagales). Syst Biol. 2012;61(2):289–313.

47. Powell CLE, Waskin S, Battistuzzi FU. Quantifying the Error of Secondary vs. Distant Primary Calibrations in a Simulated Environment. Front Genet. 2020;11(March):1–9.

48. Williams SM, McDowell JR, Bennett M, Graves JE, Ovenden JR. Analysis of whole mitochondrial genome sequences increases phylogenetic resolution of istiophorid billfishes. Bull Mar Sci. 2018;94(1):73–84.

49. Chan KMA, Levin SA. Leaky prezygotic isolation and porous genomes: Rapid introgression of maternally inherited DNA. Evolution (N Y). 2005;59(4):720–9.

50. Funk DJ, Omland KE. Species-Level Paraphyly and Polyphyly: Frequency, Causes, and Consequences, with Insights from Animal Mitochondrial DNA. Annu Rev Ecol Evol Syst. 2003;34:397–423.

51. Toews DPL, Brelsford A. The biogeography of mitochondrial and nuclear discordance in animals. Mol Ecol. 2012;21(16):3907–30.

52. Ballard JWO, Rand DM. The population biology of mitochondrial DNA and its phylogenetic implications. Annu Rev Ecol Evol Syst. 2005;36:621–42.

53. Perea S, Vukić J, Šanda R, Doadrio I. Ancient mitochondrial capture as factor promoting mitonuclear discordance in freshwater fishes: A case study in the genus *Squalius* (Actinopterygii, Cyprinidae) in Greece. PLoS One. 2016;11(12).

54. Anderson JD, Karel WJ. Genetic evidence for asymmetric hybridization between menhadens (*Brevoortia* spp.) from peninsular Florida. J Fish Biol. 2007;71(SUPPL. B):235–49.

55. Jolly MT, Maitland PS, Genner MJ. Genetic monitoring of two decades of hybridization between allis shad (*Alosa alosa*) and twaite shad (*Alosa fallax*). Conserv Genet. 2011;12(4):1087–100.

56. Hasselman DJ, Argo EE, McBride MC, Bentzen P, Schultz TF, Perez-Umphrey AA, et al. Human disturbance causes the formation of a hybrid swarm between two naturally sympatric fish species. Mol Ecol. 2014;23(5):1137–52.

57. Alexandrino P, Faria R, Linhares D, Castro F, Le Corre M, Sabatié R, et al. Interspecific differentiation and intraspecific substructure in two closely related clupeids with extensive hybridization, *Alosa alosa* and *Alosa fallax*. J Fish Biol. 2006;69(SUPPL. B):242–59.

58. Junker J, Rick JA, McIntyre PB, Kimirei I, Sweke EA, Mosille JB, et al. Structural genomic variation leads to genetic differentiation in Lake Tanganyika's sardines. Mol Ecol. 2020;29(17):3277–98.

59. Dufresnes C, Litvinchuk SN, Rozenblut-Kościsty B, Rodrigues N, Perrin N, Crochet PA, et al. Hybridization and introgression between toads with different sex chromosome systems. Evol Lett. 2020;4(5):444–56.

60. Van Sickel WA, Kominz MA, Miller KG, Browning J V. Late Cretaceous and Cenozoic sea-level estimates: Backstripping analysis of borehole data, onshore New Jersey. Basin Res. 2004;16(4):451–65.

61. Haq BU, Hardenbohl J, Vail PR. Chronology of fluctuating sea levels since the Triassic (250 million years ago to present). Science (80-). 1987;235(4):1156–67.

62. Beadle LC. The inland waters of Africa. London: Longman; 1974.

63. Peters C., O'Brien EM. Palaeo-lake Congo: implications for Africa's late Cenozoic climate—some unanswered questions. In: Palaeoecology of Africa and the Surrounding Islands, Volume 27. 2001.

64. Seehausen O, Terai Y, Magalhaes IS, Carleton KL, Mrosso HDJ, Miyagi R, et al. Speciation through sensory drive in cichlid fish. Nature. 2008;455(7213):620–6.

65. Rundle HD, Nosil P. Ecological speciation. Ecol Lett. 2005;8(3):336–52.

66. Seehausen O, Wagner CE. Speciation in Freshwater Fishes. Annu Rev Ecol Evol Syst. 2014;45:621–51.

67. Bernatchez L, Renaut S, Whiteley AR, Derome N, Jeukens J, Landry L, et al. On the origin of species: Insights from the ecological genomics of lake whitefish. Philos Trans R Soc B Biol Sci. 2010;365(1547):1783–800.

68. Mulimbwa NT, Milec LJM, Raeymaekers JAM, Sarvala J, Plisnier P, Marwa B, et al. Spatial and seasonal variation in reproductive indices of the clupeids *Limnothrissa miodon* and *Stolothrissa tanganicae* in the Congolese waters of northern Lake Tanganyika. Belgian J Zool. 2022;15:13–31.

69. Bernacsek GM, Lopes S. Cahora Bassa (Mozambique). CIFA Tech Pap. 1984;10:21–4.

70. Marshall BE. Why is *Limnothrissa miodon* such a successful introduced species and is there anywhere else we should put it? In: Pitcher TJ, Hart PJB, editors. The Impact of Species Changes in African Lakes. 1995. p. 527–45.

71. Kullander SO, Roberts TR. Out of Lake Tanganyika: Endemic Lake fishes inhabit rapids of the Lukuga River. Ichthyol Explor Freshwaters. 2011;22(4):355–76.

72. Tajima F, Nei M. Genetic drift and estimation of effective population size. Genetics. 1981;98:625–40.

73. Kmentová N, Koblmüller S, Van Steenberge M, Raeymaekers JAM, Artois T, De Keyzer ELR, et al. Weak population structure and recent demographic expansion of the monogenean parasite *Kapentagyrus* spp. infecting clupeid fishes of Lake Tanganyika, East Africa. Int J Parasitol. 2020;50(6–7):471–86.

74. Koblmüller S, Zangl L, Börger C, Daill D, Vanhove MPM, Sturmbauer C, et al. Only true pelagics mix: comparative phylogeography of deepwater bathybatine cichlids from Lake Tanganyika. Hydrobiologia. 2018;832(1):93–103.

75. Settle ML. Proc10xG. 2017. p. https://github.com/ucdavis-bioinformatics/proc10x.

76. Andrews S, others. FastQC: a quality control tool for high throughput sequence data. 2010. Https://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/. 2010. p. http://www.bioinformatics.babraham.ac.uk/projects/.

77. Meng G, Li Y, Yang C, Liu S. MitoZ: a toolkit for animal mitochondrial genome assembly, annotation and visualization. Nucleic Acids Res. Oxford University Press; 2019;(29):1–7.

78. Iwasaki W, Fukunaga T, Isagozawa R, Yamada K, Maeda Y, Satoh TP, et al. Mitofish and mitoannotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol Biol Evol. 2013;30(11):2531–40.

79. Kans J. E-utilities on the Unix Command Line. In: Entrez Programming Utilities Help. Bethesda (MD): National Center for Biotechnology Information (US); 2013.

80. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. Mega X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol. 2018;35:1547–9.

81. Abascal F, Zardoya R, Telford MJ. TranslatorX: Multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res. 2010;38(SUPPL. 2):7–13.

82. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.

83. Borowiec ML. AMAS: A fast tool for alignment manipulation and computing of summary statistics. PeerJ. 2016;2016(1).

84. Strimmer K, Von Haeseler A. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci U S A. 1997;94(13):6815–9.

85. Schmidt HA, Strimmer K, Vingron M, Von Haeseler A. TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics. 2002;18(3):502–4.

86. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and high-performance computing Europe PMC Funders Group. Nat Methods. 2012;9(8):772.

87. Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, et al. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34(12):3299–302.

88. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol. 2021;38(7):3022–7.

89. R core team. R: A language and environment for statistical computing [Internet]. Vienna, Austria; 2020. Available from: https://www.r-project.org/

90. Bernt M, Merkle D, Ramsch K, Fritzsch G, Perseke M, Bernhard D, et al. CREx: Inferring genomic rearrangements based on common intervals. Bioinformatics. 2007;23(21):2957–8.

91. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.

92. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587–9.

93. Minh BQ, Nguyen MAT, Von Haeseler A. Ultrafast approximation for phylogenetic bootstrap. Mol Biol Evol. 2013;30(5):1188–95.

94. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307–21.

95. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003;19(12):1572–4.

96. Gelman A, Rubin D. Inference from Iterative Simulation using Multiple Sequences. Stat Sci. 1992;7(4):457–72.

97. Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, et al. BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. PLoS Comput Biol. 2019;15(4):1–28.

98. Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 2018;67(5):901–4.
99. Rambaut A. FigTree. 2018. p. http://tree.bio.ed.ac.uk/software/figtree/.
100. Lavoué S, Bertrand JAM, Wang HY, Chen WJ, Ho HC, Motomura H, et al. Molecular systematics of the anchovy genus *Encrasicholina* in the Northwest Pacific. PLoS One. 2017;12(7):1–16.
101. Malabarba MC, Di Dario F. A new predatory herring-like fish (Teleostei: Clupeiformes) from the early Cretaceous of Brazil, and implications for relationships in the Clupeoidei. Zool J Linn Soc. 2017;180(1):175–94.
102. Marramà G, Carnevale G. An Eocene anchovy from Monte Bolca, Italy: The earliest known record for the family Engraulidae. Geol Mag. 2016;153(1):84–94.
103. Miller RR. First Fossil Record (Plio-Pleistocene) of Threadfin Shad, *Dorosoma petenense*, from the Gatuña Formation of Southeastern New Mexico. J Paleontelogy. 1982;56(2):423–5.
104. Rohatgi A. WebPlotDigitizer [Internet]. Pacifica, California, USA; 2021. Available from: https://automeris.io/WebPlotDigitizer Additional file information Additional file 1.xls: Taxonomic information, accession numbers and references of mitochondrial genomes used for phylogenetic analyses.

Figures

Figure 1

Mitochondrial genomes of *Stolothrissa tanganicae* (upper) and *Limnothrissa miodon* (lower). Inner circle shading indicates GC-content. Outer circle: black regions are protein coding genes, red regions are tRNA genes, beige regions are rRNA genes, and brown is the D-loop. Regions closer to the centre are located on the minus strand.

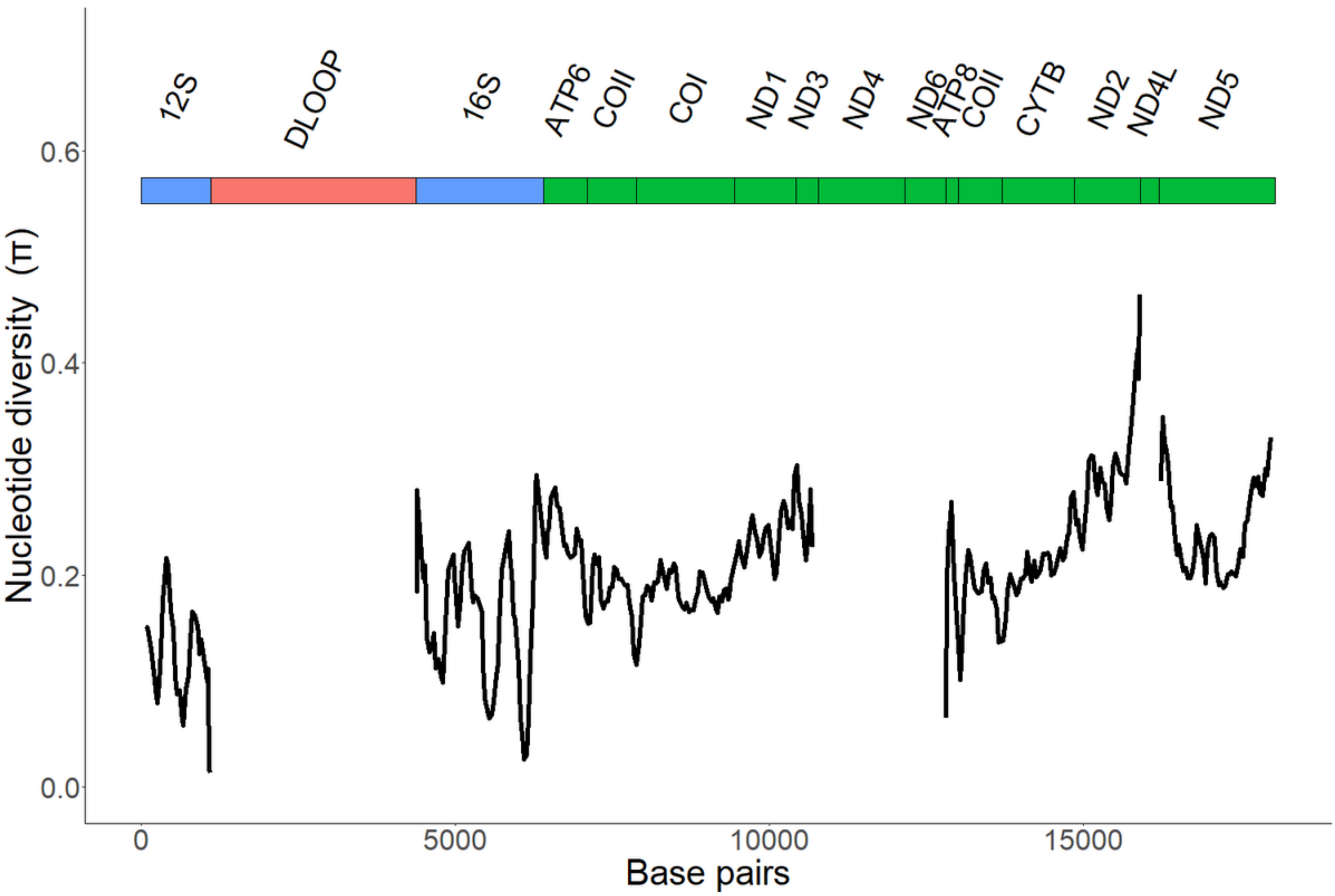

Figure 2

Nucleotide diversity (π) at mitochondrial PCGs (green) and rRNA genes (blue). π was calculated for 109 clupeiform and 6 non-clupeiform fish species using a sliding window with a window size of 150 bp and steps of 35 bp. Values could not be calculated for the gene coding for ND4, ND6, ND4L and for the D-loop (red).

Figure 3

Outgroup-rooted maximum likelihood phylogeny of Clupeiformes. Topology and branch lengths were estimated based on mitochondrial protein-coding genes, rRNA genes and D-loop sequence of 109 clupeiforms and 6 non-clupeiforms. Node support was assessed by Shimodaira-Hasegawa-like approximate likelihood ratio tests (SH-aLRT%) and ultrafast bootstrap (UFBoot%). Nodes with SH-aLRT% < 75 and UFBoot% < 90 were polytomized and their support values are not shown. The scale bar indicates evolutionary distance (expected number of nucleotide substitutions per site). Subfamilies are indicated on the right side in black, families in colour. The tribe Pellonulini is highlighted in green.

Figure 4

Outgroup-rooted Bayesian phylogeny of Clupeiformes. Topology and branch lengths were estimated based on mitochondrial protein-coding genes, rRNA genes and D-loop sequence of 109 clupeiforms and 6 non-clupeiforms. Node support was assessed by Bayesian posterior probabilities (BPP). Nodes with BPP < 0.85 were polytomized and their support values are not shown. Probabilities were rounded to the nearest 0.01. The scale bar indicates evolutionary distance (expected number of nucleotide substitutions per site). Subfamilies are indicated on the right side in black, families in colour. The tribe Pellonulini is highlighted in green.

Figure 5

Outgroup-rooted time-calibrated phylogeny of Clupeiformes. Divergence times were estimated using BEAST, based on the first and second codon positions of 13 mitochondrial protein coding genes of 109 clupeiform and 6 non-clupeiform fishes. Blue bars represent Bayesian 95% credibility intervals. Calibration points (C1-C3) and fossil validation points (F1-F3) are indicated on the corresponding nodes. The tribe Pellonulini is highlighted in green.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.xls