Molecular Phylogenetics and Evolution 124 (2018) 172–180

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Multilocus phylogenetic analysis of the first molecular data from the rare and monotypic Amarsipidae places the family within the Pelagia and highlights limitations of existing data sets in resolving pelagian interrelationships

Matthew A. Campbell[a,b,*], Tetsuya Sado[c], Chuya Shinzato[d], Ryo Koyanagi[e], Makoto Okamoto[f], Masaki Miya[c,*]

[a] Department of Ecology and Evolutionary Biology, University of California Santa Cruz, Santa Cruz, CA 95064 USA
[b] Fisheries Ecology Division, Southwest Fisheries Science Center, National Marine Fisheries Service, Santa Cruz, CA 95060 USA
[c] Department of Ecology and Environmental Sciences, Natural History Museum and Institute, Chiba 260-8682, Japan
[d] Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan
[e] DNA Sequencing Section, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan
[f] Marine Fisheries Research and Development Center (JAMARC), Fisheries Research Agency, Kanagawa 220-6115, Japan

ARTICLE INFO

Keywords:
Amarsipa
Amarsipus carlsbergi
Pelagiaria
Scombriformes
Scombrimorpharia

ABSTRACT

The Pelagia is a recently delineated group of fishes, comprising fifteen families formerly placed in six perciform suborders. The Pelagia was lately recognized as it encompasses huge morphological diversity and only in the last few years have large-scale molecular phylogenetic studies been undertaken that could unite such morphologically disparate lineages. Due to the recent erection of Pelagia, the composition of the taxon is not entirely certain. Five families of the former perciform suborder Stromateoidei have been identified as pelagians. However, the sixth stromateoid subfamily Amarsipidae is a rare monotypic family that has distinctive meristic and morphological characteristics from that of other stromateoids such as the lack of a pharyngeal sac. We examine molecular data generated from the sole species in Amarsipidae, *Amarsipus carlsbergi*, and demonstrate that it is clearly nested within Pelagia. As with two previous studies that have the breadth of sampling to evaluate pelagian intra-relationships, we find high support for monophyly of most family-level taxonomic units but statistical support for early-branching nodes in the pelagian tree is very low. We conduct the first analyses of Pelagia incorporating the multispecies coalescent and are limited by a high degree of missing loci, or, incomplete taxon sampling. The high degree of missing data across a complete sampling of pelagian lineages along with the deep time scale and rapid radiation of the lineage contribute to poor resolution of early-branching relationships within Pelagia that cannot be resolved with current data sets. Currently available data are either mitochondrial genomes or a super matrix of relatively few loci with a high degree of missing data. A new and independent dataset of numerous phylogenetic loci derived from high-throughput sequencing technology may reduce uncertainty within the Pelagia and provide insights into this adaptive radiation.

1. Introduction

The ocean represents an immense habitat and a plethora of potential niches. The upper 200 m of the ocean, the epipelagic zone, accounts for 70% of the earth's surface with spatial and temporal variation in primary production and consequently food availability (Allen and Cross, 2006; Helfman et al., 1997). Mobility is a key feature of epipelagic fishes to accommodate spatial and temporal variation in food supply, and despite the outward uniformity of the open ocean, fishes of the epipelagic zone have a diversity of morphologies to utilize a broad set of niches. A taxon of fishes occupying the pelagic realm, the Pelagia (*sensu* Miya et al., 2013), consists of fishes of disparate morphologies formerly found in six perciform suborders (Icosteoidei, Percoidei, Scombroidei, Scombrolabracoidei, Stromateoidei, and Trachinioidei) and fifteen families. The pelagians originate from an adaptive radiation occurring as a result of ecological openings in the epipelagic following the Cretaceous-Tertiary mass extinction (Miya et al., 2013). The Cretaceous-Tertiary mass extinction removed previous large fishes

* Corresponding authors at: Research Faculty of Agriculture, Hokkaido University, Sapporo 060-8589, Japan (M.A. Campbell).
E-mail addresses: drmaccampbell@gmail.com (M.A. Campbell), miya@chiba-muse.or.jp (M. Miya).

https://doi.org/10.1016/j.ympev.2018.03.008
Received 26 July 2017; Received in revised form 31 January 2018; Accepted 7 March 2018
Available online 08 March 2018

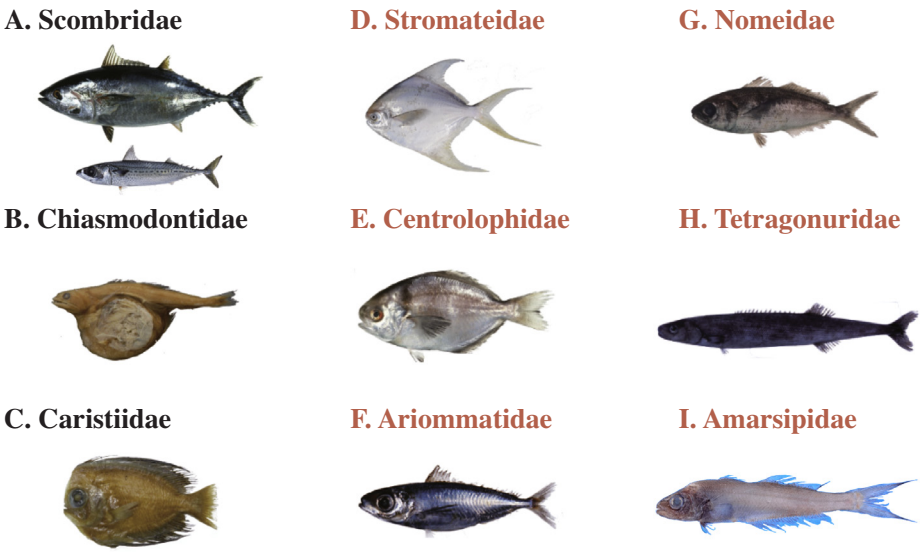

Fig. 1. Representative images of pelagians. Three non-stromateoid families are depicted: 1A Scombridae; 1B Chiasmodontidae; 1C Caristiidae. (D–I) Labeled in red are stromateoid families with (I) Amarsipidae. (A–H) are adapted from Miya et al. (2013). (I) is a picture of the specimen examined for this manuscript as described in the methods. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

occupying pelagic niches through the collapse of the oceanic food chain and allowed the rapid radiation of Pelagia lineage into now vacant niches developing diverse morphologies from a single common ancestor as they did so (Cavin, 2002; Friedman, 2009, 2010; Friedman and Sallan, 2012; Miya et al., 2013).

Monophyly of morphologically disparate pelagian lineages has been found in independent studies aside from Miya et al. (2013) and the group is also known as the Pelagiaria (Betancur, 2013; Betancur et al., 2016, 2017; Chen et al., 2003; Sanciangco et al., 2016). This diverse assemblage contains tunas and their relatives (Scombridae) that are potentially large species, such as *Thunnus thynnus* which commonly grows to 2.0 m in length (Collette and Nauen, 1983). Scombrids are adapted to sustained and rapid swimming with robust bodies (Magnuson, 1979) (Fig. 1A). The snaketooth fishes (Chiasmodontidae) and manefishes (Caristiidae) are pelagians that have alternative lifestyles and morphologies to scombrids, but are nonetheless close relatives that occupy pelagic waters (Fig. 1B and C). Molecular phylogenetic analysis has shown five of the six families of Stromateoidei (Fig. 1D–H) to be placed with the Pelagia (Miya et al., 2013); however, the family Amarsipidae (Fig. 1I) previously has not been studied in a molecular phylogenetic framework and has not been represented in molecular studies.

The Amarsipidae is rare and monotypic, with *Amarsipus carlsbergi* (the amarsipa) being first described in 1969 from 47 young collected in the open ocean between 8 and 68 mm in standard length (Haedrich, 1969). Morphologically *Amarsipus* was considered distinct from other stromateoids by the absence of a key feature of this group, the pharyngeal sac, a toothed pouch found in the digestive tract posterior to the last gill arch. Additionally *Amarsipus* has several meristic differences and fang-like teeth on pharyngeal bones that distinguish it from other stromateoids (Haedrich, 1969; Konovalenko and Piotrovskiy, 1989; Okamoto et al., 2011). Because of the morphological distinctiveness of *Amarsipus,* it was assigned to a new family when first described (Haedrich, 1969). Adult *Amarsipus* were first documented in 1988 from collections targeting sound scattering layers at night, indicating *Amarsipus* undergoes diel migrations and associates with other fishes of the deep scattering layer (Konovalenko and Piotrovskiy, 1989). A single adult individual was captured from Vietnamese waters from between 760 and 800 m (Konovalenko and Piotrovskiy, 1989) though 12–130 m in depth is more typical (Haedrich, 1969; Konovalenko and Piotrovskiy, 1989; Parin and Piotrovsky, 2004). All specimens to date have been captured at night, and although the fish is uncommon, it has a wide distribution across marine waters of the Indo-Pacific in tropical latitudes (Haedrich, 1969; Konovalenko and Piotrovskiy, 1989; Parin and Piotrovsky, 2004). In general, the ecology of *Amarsipus* remains largely unknown. To further refine the composition of Pelagia and to understand this radiation of fishes into the open sea, we examine DNA sequence data to place *Amarsipus* in the fish tree of life.

2. Methods

2.1. Sample collection, DNA extraction, and sequencing

A tissue sample was taken from a single individual of *Amarsipus carlsbergi* (96.7 mm SL; CBM-ZF 17750) collected at the Nagasaki Fish Market (Nagasaki City, Nagasaki Prefecture, Japan). Total genomic DNA was extracted with a Qiagen DNeasy Blood & Tissue Kit (Qiagen, Inc., www.qiagen.com) according to the manufacturer's instructions. Extracted DNA was sequenced by preparing 300 ng of DNA with a KAPA HyperPlus Kit (KAPA Biosystems Inc., www.kapabiosystems.com) and prepared libraries were sequenced with an Illumina HiSeq 2500 with 150 base pair (bp) paired-end sequencing and version 3 chemistry.

2.2. DNA sequence data quality control and assembly

Low quality bases were removed and adapter contamination removed with Trimmomatic version 0.36 (Bolger et al., 2014) with the following options: ILLUMINACLIP 2:30:10, LEADING 3, TRAILING 3, SLIDINGWINDOW 4:15, MINLEN 36. Sequence quality was evaluated after trimming and adapter removal with FastQC version 0.11.15 (Andrews, 2010) to verify the quality of sequences and removal of adapter sequences. Overlapping paired-end reads were merged with FLASH version 1.2.11 (Magoč and Salzberg, 2011) using default settings except maximum overlap was increased (-M 140).

Assembly of surviving paired-end reads was conducted with Genome Region Assembly by Baiting (GRAbB, downloaded February 8, 2017) (Brankovics et al., 2016). GRAbB drives various programs to first find reads matching target sequences, then a *de novo* assembly is conducted and completion checked. For target sequences we obtained the most complete set of loci from a pelagian in Betancur et al. (2013a), *Peprilus simillimus,* then merged data from two other species, *Sarda sarda* and *Scomberomorus regalis*, to create a complete matrix of the twenty-one loci in Betancur et al. (2013a) (Table 1). Matching of reads to target sequences "baiting" was done with mirabait part of MIRA version 4.0 (Chevreux et al., 1999), assembly of baited sequences was conducted with Velvet version 1.2.10 (Zerbino and Birney, 2008) and completion was checked with Exonerate version 2.2.0 (Slater and

Table 1
Gene abbreviation, reference accession (if used), and species of origin of nucleotide sequences used as baits. If the targeted gene was determined to be assembled, coverage from Velvet output, the length of the assembled gene and GenBank accession are given. For all alignments the total alignment length is reported.

Gene	Reference accession	Reference species	Assembly of target gene?	Measure of coverage	Gene length for alignment (base pairs)	Overall alignment length	Amarsipidae GenBank accession
ENC1	JX188994.1	*Sarda sarda*	No			657	
FICD	KC825745.1	*Peprilus simillimus*	Yes	36.98	541	699	MH043138
GLYT	JX188822.1	*Sarda sarda*	Yes	31.85	825	864	MH043139
Hoxc6a	KC826172.1	*Peprilus simillimus*	Yes	32.68	482	736	MH043140
KIAA1239	JQ939828.1	*Peprilus simillimus*	Yes	29.61	771	918	MH043141
MYH6	JQ939530.1	*Peprilus simillimus*	No			744	
PANX2	KC827815.1	*Peprilus simillimus*	Yes	29.58	521	753	MH043142
PLAGL2	JQ937581.1	*Peprilus simillimus*	No			780	
PTCHD1	KC828508.1	*Peprilus simillimus*	Yes	27.81	470	750	MH043143
RAG1	JX189923.1	*Sarda sarda*	Yes	31.31	1110	1464	MH043144
RAG2	DQ874766.1	*Scomberomorus regalis*	Yes	26.64	549	1128	MH043145
RHOD	DQ874798.1	*Scomberomorus regalis*	No			852	
RIPK4	KC829136.1	*Peprilus simillimus*	Yes	20.05	533	645	MH043146
SH3PX3	JQ940145.1	*Peprilus simillimus*	No			705	
SIDKEY	JQ937715.1	*Peprilus simillimus*	No			1101	
SREB2	JX190061.1	*Sarda sarda*	Yes	26.45	581	987	MH043147
SVEP1	KC830159.1	*Scomberomorus regalis*	Yes	32.12	741	808	MH043148
TBR1	JX189299.1	*Sarda sarda*	No			681	
VCPIP	KC830658.1	*Sarda sarda*	No			765	
ZIC1	KC831070.1	*Peprilus simillimus*	Yes	28.45	441	855	MH043149
16S	JQ938996.1	*Peprilus simillimus*	No				
16S	*de novo*	NA	Yes	18.22	1691	1851	AP018529

Birney, 2005). Assembled sequences were verified through BLASTN version 2.4.0 (Altschul et al., 1997; Morgulis et al., 2008) by comparison to the original target sequence and the NCBI nucleotide reference database. Assemblies with paralogous sequences, that is those with two or more near identical BLASTN matches, were not retained for further analysis. The 16S mitochondrial ribosomal DNA (16S) did not assemble into a contiguous piece with this approach.

To assemble the complete *Amarsipus* mitogenome, trimmed and unmerged paired-end reads were mapped to seventy-two pelagian mitogenomes with Bowtie2 version 2.2.9 (Langmead and Salzberg, 2012) with default settings. Concordantly mapped reads, that is with both reads mapping, and pairs of reads with a single read mapping were extracted with SAMtools version 1.3.1 (Li et al., 2009) from the resulting alignments. The selected mapping read pairs were normalized to a target depth of 35 with bbnorm, part of BBTools version 36.86 (http://jgi.doe.gov/data-and-tools/bbtools/). Normalized reads were assembled across kmer values of 67 to 95 with VelvetOptimiser version 2.2.5 (Gladman and Seemann, 2012). A mitochondrial contiguous sequence (contig) was identified from the resulting assembly by comparison to the NCBI nucleotide reference database with BLASTN (Altschul et al., 1997; Morgulis et al., 2008) and annotated with mitoannotator through the MitoFish website (http://mitofish.aori.u-tokyo.ac.jp/, Accessed February 2, 2017) (Iwasaki et al., 2013).

2.3. Sequence alignment

Sequence data from all the Scombrimorpharia from Betancur et al. (2013a) was downloaded from Dryad (Betancur, 2013). In addition, recently published nuclear protein coding gene data from Arripidae (Sanciangco et al., 2016) were obtained from GenBank (KT883742; KT883770; KT883818; KT883832) and combined with mitochondrial data (KR153522) (Supplemental Table S1). To represent Tetragonuridae, which was not present in Betancur et al. (2013a), 16S sequence data was obtained from mitochondrial genomes (Supplemental Table S1). All sequences were re-aligned with MAFFT version 7.130b (Katoh et al., 2002; Katoh and Toh, 2008). Long ends of *Amarsipus* derived sequences were trimmed and complete translation into amino acids for protein coding sequences was checked with Mesquite version 3.04 (Maddison and Maddison, 2011). Two data treatments were generated, where all codon positions were included and not recoded ($1_N2_N3_N$) and where third codon positions were recoded to purines or pyrimidines (RY-coding, $1_N2_N3_{RY}$). The resulting data matrices contained many taxa with few genes sequenced; therefore we created "Subset" data matrices with each taxon represented by 10 or more genes in the alignment. A total of four data matrices were created in total: All Data $1_N2_N3_N$, All Data $1_N2_N3_{RY}$, Subset $1_N2_N3_N$, and Subset $1_N2_N3_{RY}$.

2.4. Concatenated phylogenetic analysis

For all four data matrices, Maximum Likelihood (ML) phylogenies were generated with Randomized A(x)ccelerated Maximum Likelihood (RAxML) version 8.2.10 (Stamatakis, 2006; Stamatakis and Ott, 2008). Confidence at nodes of the inferred phylogenies was measured through the rapid bootstrap algorithm (-f a) with automatic stopping specified (-N autoMRE) in RAxML with partitions modeled under the General Time Reversible (GTR) model of nucleotide evolution with Gamma distributed rate variation (Γ). Partitioning of data is a strategy for incorporating heterogeneity among sites and can lead to substantial and important changes of topologies in phylogenetic analyses e.g. (Blair and Murphy, 2011; Campbell et al., 2014b; Nylander et al., 2004). Importantly, the partitioning strategy for a phylogenetic analysis must be defined *a priori*, therefore the four data matrices (All Data $1_N2_N3_N$, All Data $1_N2_N3_{RY}$, Subset $1_N2_N3_N$, Subset $1_N2_N3_{RY}$) were partitioned by gene under the assumption that the behavior of each genes is independent. Next, we created objectively defined partitioning strategies with PartitionFinder version 2.1.1 (Lanfear et al., 2012). We evaluated possible combinations of the 16S data and each codon position of each gene to find the preferred partitioning strategy for each data set for analysis with RAxML as identified by the Bayesian information criterion (BIC). The specified options for PartitionFinder were branchlengths = linked, models = GTR + G, model_selection = BIC, and search = greedy.

2.5. Coalescent phylogenetic analyses

The Pelagia is an adaptive radiation of fishes; therefore, Incomplete Lineage Sorting (ILS) is present and should be accounted for in analyses through methodology incorporating the multispecies coalescent (Knowles and Kubatko, 2011). Additionally, in concatenated analyses

small regions of alignments have been indicated to be overly influential and can mislead analyses (Shen et al., 2017). To investigate the influence of both of these potential sources of error we conducted summary coalescent analysis of the All Data $1_N2_N3_N$ and Subset $1_N2_N3_N$ data matrices. For each gene in the alignment for both of these data matrices, a ML phylogeny was generated with RAxML version 8.0.19 with a GTR + Γ model of nucleotide evolution. The resulting best trees of each gene from RAxML were analyzed with the Accurate Species TRee ALgorithm (ASTRAL) III with version 5.5.9 of the program code to produce a species tree (Mirarab et al., 2014; Sayyari and Mirarab, 2016; Zhang et al., 2017).

3. Results and discussion

3.1. DNA sequence data quality control and assembly

332,856,113 raw read pairs were processed by Trimmomatic resulting in 93.29% surviving as paired reads. Of these, 58% (181,354,946) are combined by FLASH resulting in 129,160,221 uncombined read pairs passing initial quality control. The GRAbB approach assembled twelve protein-coding gene loci after verification of homology and assembly of the target gene through BLASTN to reference sequences of entire assemblies (Table 1). Subsequent *de novo* assembly of mitochondrial mapping read pairs produced a single 16,404 bp contig from which 1691 bp are annotated by MitoAnnotator to be 16S (Table 1).

3.2. Sequence alignment

A total of 9256 nucleotides were aligned from Amarsipidae assemblies into the larger data matrix. From this total, 7565 characters are from twelve nuclear protein-coding genes. The All Data $1_N2_N3_N$ and All Data $1_N2_N3_{RY}$ matrices are 18,742 characters in length and have 65.18% missing data and gaps with 7616 (All Data $1_N2_N3_N$) and 6091 (All Data $1_N2_N3_{RY}$) distinct alignment patterns. With the All Data matrices, 74 taxa are represented with 46 previously identified pelagians. Restricting the data matrices to taxa with at least 10 gene sequences results in 31 taxa for analysis, with 22 previously identified pelagians. The Subset $1_N2_N3_N$ and Subset $1_N2_N3_{RY}$ matrices are 18,647 characters in length with 51.08% gaps or missing data and have 5633 distinct alignment patterns (Subset $1_N2_N3_N$) and 4654 distinct alignment patterns (Subset $1_N2_N3_{RY}$). The All Data $1_N2_N3_N$ alignment is available in the Data Supplement.

3.3. Phylogenetic analysis

Amarsipidae is nested deep within the Pelagia most closely related to another stromateoid family, Tetragonuridae, with low support values. Overall, changes to the alignments (recoding) and partitioning (by gene or by PartitionFinder) create substantial changes in tree topologies. The All Data $1_N2_N3_N$ ML phylogenetic analysis partitioned by gene is presented in Fig. 2 and the All Data $1_N2_N3_{RY}$ partitioned by gene phylogenetic tree is presented as Fig. 3. Partitioning as determined by PartitionFinder is described in the data supplement, with ML trees from All Data $1_N2_N3_N$ ML and All Data $1_N2_N3_{RY}$ presented as Supplemental Figs. S1 and S2. Both tree files as well as those with partitioning schemes created by PartitionFinder are available in the Data Supplement. The bootstrap support (BS) for a sister Amarsipidae and Tetragonuridae relationship is 66% in the All Data $1_N2_N3_N$ analysis and BS = 60% in the All Data $1_N2_N3_{RY}$ analysis when partitioned by gene. In the All Data $1_N2_N3_N$ and All Data $1_N2_N3_{RY}$ analyses partitioned by gene, the other four stromateoid families excluding Amarsipidae and Tetragonuridae form a possible paraphyletic assemblage with very low support, BS = 22% and BS = 26% respectively. In general, family level taxonomic units such as the Scombridae and Trichiuridae are well supported. The Scombridae is supported with a BS = 99% with both All Data analyses that are partitioned by gene, while the Trichiuridae in both analyses is highly supported as well (BS = 100%). The Gempylidae is these analyses is only moderately supported, with a BS = 81% in the $1_N2_N3_N$ analysis and BS = 83% in the $1_N2_N3_{RY}$ analysis.

The Subset $1_N2_N3_N$ ML phylogenetic analysis partitioned by gene places Amarsipidae within the Pelagia and sister (BS = 46%) to a monophyletic Scombridae (BS = 100%) (Fig. 4A). Recoding of the third codon positions to purines and pyrimidines (Subset $1_N2_N3_{RY}$) is presented as Supplemental Fig. S3. Summary coalescent analysis confirms the monophyly of Pelagian taxa in this analysis with high posterior probability (pp) of 1.00 (Fig. 4B, All Data $1_N2_N3_N$ for comparison is included as Supplemental Fig. S4). Summary coalescent analysis supports that Amarsipidae is a near relative of Scombridae, with a sister Amarsipidae and Scombridae receiving moderate support (pp = 0.77). The monophyly of Scombridae receives low support (pp = 0.50) in the Subset summary coalescence analysis. All tree files from the subset analyses are available in the Data Supplement.

3.4. Higher level relationships within Pelagia remain elusive

The phylogenetic analyses presented in this manuscript clearly place Amarsipidae within the Pelagia (Figs. 2–4, Supplemental Figs. S1–S4). However, further interpretations of the relationship of Amarsipidae to other pelagians are very limited. Amarsipidae receives some support as the sister lineage of Tetragonuridae, but it is important to consider the representation of Amarsipidae with characters across thirteen genetic loci is much greater than that of Tetragonuridae, which only is represented by mitochondrial 16S ribosomal DNA data. Between the lower statistical support value for the Amarsipidae + Tetragonuridae sister relationship (BS = 60–68%) and the limited sampling of characters for Tetragonuridae it is not clear what the closest pelagian relative of Amarsipidae is solely based on the molecular evidence presented in this study. A cladistic analysis of a limited number of characters (27, six of which are related to the pharyngeal sac and papillae) indicated that Amarsipidae was sister to all other stromateoids (Horn, 1984), consequently the pharyngeal sac would not be a synapomorphy of Stromateoidei. Subsequently, Doiuchi et al. (2004) examined the morphological relationships of stromateoid fishes through cladistic analysis of 43 characters. Centrolophids were found to be non-monophyletic, and an unresolved clade composed of Amarsipidae, Tetragonuridae, Ariommatidae, Nomeidae and Stromateidae is supported (Doiuchi et al., 2004), and the absence of a pharyngeal sac in Amarsipidae is a reversal. The molecular data provides some refinement of these relationships, as a close relationship between Amarsipidae and Tetragonuridae is supportable within the pelagia. Following Doiuchi et al. (2004), the absence of the pharyngeal sac in Amarsipidae may be a reversal if stromateoid monophyly is assumed.

Above the family level, nodes are very weakly supported in the analyses (Figs. 2–4, Supplemental Figs. S1–S4). Nonetheless, there are notable similarities and differences between previous studies and this one that are insightful. This study builds off the existing multilocus data matrix of Betancur et al. (2013a), in which the Scombridae is not monophyletic. The Scombridae is highly supported by the mitogenomic dataset of Miya et al. (2013) (BS = 97%). Our realignment of the Betancur et al. (2013a) data matrix, additional sampling of three pelagian lineages, and different analyses result in a monophyletic Scombridae that is strongly supported across concatenated analyses (BS = 98–100%) and present, though not necessarily highly supported in summary coalescence analyses, All Data $1_N2_N3_N$ pp = 0.88, Subset $1_N2_N3_N$ pp = 0.49.

We also examined the Arripidae, a family that Miya et al. (2013) had placed as sister to the Pomatomidae with mitogenomic data. Sanciangco et al. (2016) also examined the evolutionary affinities of the Arripidae through the sampling of Arripidae and incorporation in the Betancur et al. (2013a) data matrix. Previously with multilocus data, Arripidae is placed with < 75% BS as the sister lineage of all other

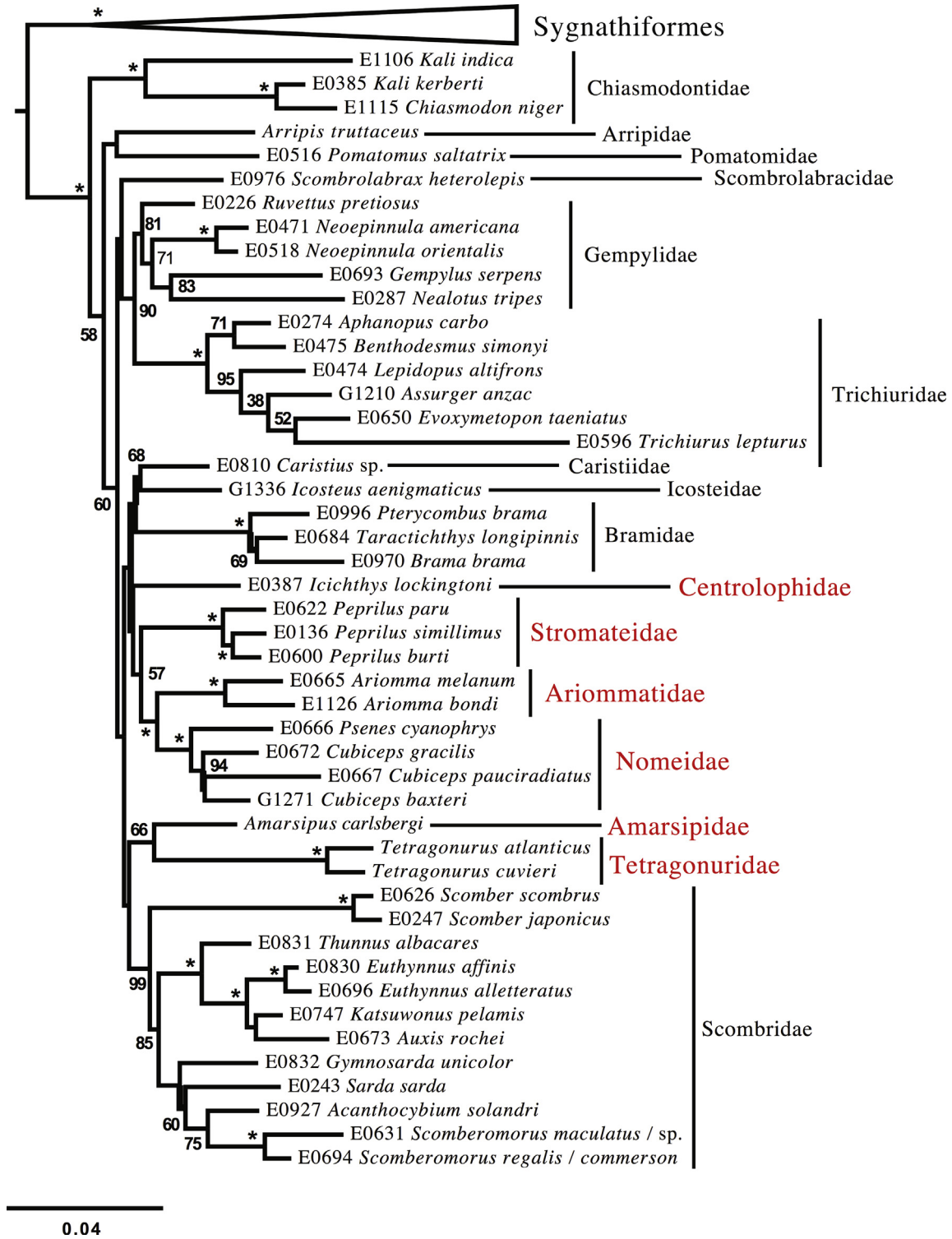

Fig. 2. A Maximum-Likelihood (ML) phylogenetic hypothesis of all previously known pelagian families and Amarsipidae. Amarsipidae is clearly placed within the Pelagia. Stromateoid families are labeled in red. The ML tree was generated from partitioned analysis of 21 genetic loci. Each locus was modeled under the General Time Reversible (GTR) model of sequence evolution with Gamma distributed rate variation (Γ). Bootstrap support values are indicated at nodes if ≥50 and by an asterisk (*) if equal to 100. The tree is rooted by sygnathiform outgroup species. Species from the Euteleost Tree of Life (EToL) alignment of Betancur et al. (2013a) are prepended with EToL identifiers. The two *Scomberomus* in this data set, E0631 and E0694, have two species epithets, as they are composed of sequences from two separate species each merged to create a genus-level composite taxon. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pelagians excluding Chiasmodontidae (Sanciangco et al., 2016). We find a weakly supported Arripidae and Pomatomidae sister relationship (BS = 48% and 44%) in our All Data $1_N2_N3_N$ and All Data $1_N2_N3_{RY}$ concatenated analyses partitioned by gene and the same relationship is strongly supported by mitogenomic data (BS = 96%).

The Gempylidae that is clearly divided into two families by Miya et al. (2013) remains monophyletic in this study. Resolution of the monophyly or not of Gempylidae with the Betancur et al. (2013a) dataset as a starting point requires representation of additional divergent lineages of Gempylidae. Gempylidae I (Miya et al., 2013) is represented

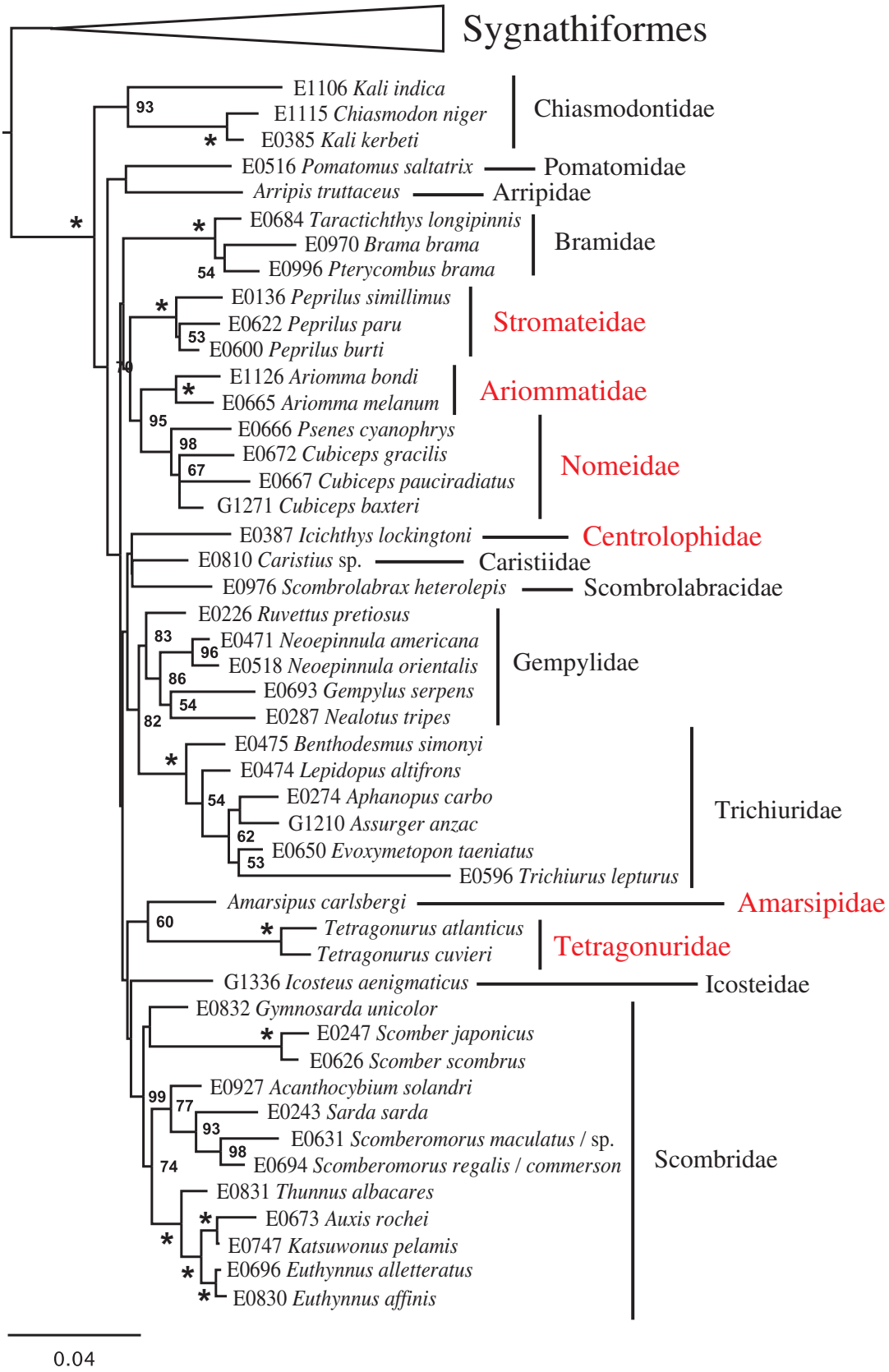

Fig. 3. A Maximum-Likelihood (ML) phylogenetic hypothesis of all previously known pelagian families and Amarsipidae. Amarsipidae is clearly placed within the Pelagia. Stromateoid families are labeled in red. The ML tree was generated from partitioned analysis of 21 genetic loci. The third codon positions of protein-coding genes were recoded as purines or pyrimidines – RY coding. Each locus was modeled under the General Time Reversible (GTR) model of sequence evolution with Gamma distributed rate variation (Γ). Bootstrap support values are indicated at nodes if ≥50 and by an asterisk (*) if equal to 100. The tree is rooted by sygnathiform outgroup species. Species from the Euteleost Tree of Life (EToL) alignment of Betancur et al. (2013a) are prepended with EToL identifiers. The two *Scomberomus* in this data set, E0631 and E0694, have two species epithets, as they are composed of sequences from two separate species each merged to create a genus-level composite taxon. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

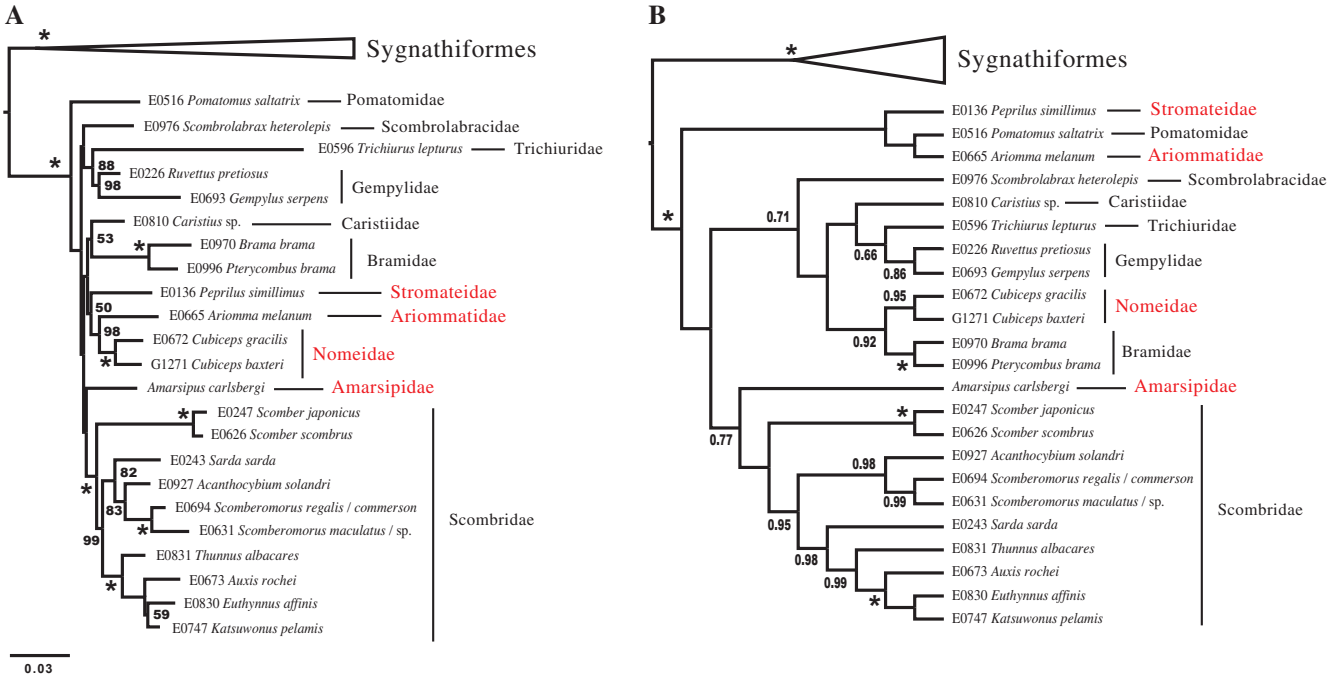

Fig. 4. Phylogenetic hypotheses of Pelagia generated from taxa with at least 10 genes sequenced in the data matrix compiled for this study, the Subset $1_N2_N3_N$ data set. (A) A Maximum-Likelihood (ML) phylogenetic tree from by-gene partitioning of 21 genetic loci modeled with the General Time Reversible (GTR) model of nucleotide evolution with Gamma distributed rate variation (Γ). Bootstrap support values are indicated at nodes if ≥50 and by an asterisk (*) if equal to 100. (B) Species tree presented as a cladogram generated by the ASTRAL III algorithm applied to 21 individually generated gene trees. A ML tree was generated for each gene tree under the GTR + Γ model of nucleotide evolution. Posterior probability values are shown at nodes when > 0.50, and an asterisk is used when the posterior probability is equal to 1.00. Both phylogenetic hypotheses support Amarsipidae as a pelagian. Each tree is rooted by sygnathiform outgroup species and the Stromateoid families are colored red. Species from the Euteleost Tree of Life (EToL) alignment of Betancur et al. (2013a) are prepended with EToL identifiers. The two *Scomberomus* in this data set, E0631 and E0694, have two species epithets, as they are composed of sequences from two separate species each merged to create a genus-level composite taxon. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

by *Gempylus serpens*, *Neopinnula americana*, *Neopinnula orientalus* and *Nealotus tripes* in this study while Gempylidae II is represented by *Ruvettus pretiosus*. To verify the non-monophyly of Gempylidae II, potential Gempylidae II lineages such as *Epinnula magistralis* and *Lepidocybium flavobrunneum* (Miya et al., 2013) should also be examined with non-mitochondrial genetic loci, as the limited sampling here does not permit a conclusion regarding gempylid monophyly.

4. Conclusions

The Pelagia tree is typified by very low support values within early-branching lineages and inconsistent placement of higher taxonomic units in this and other studies (Betancur et al., 2013a; Miya et al., 2013; Sanciangco et al., 2016). The inclusion of Amarsipidae is an important step in resolving the higher-level relationships of Pelagia. Increasing sampling of lineages is a strategy for inferring correct relationships (Chen and Mayden, 2010; Hillis, 1998; Zwickl and Hillis, 2002), as well as decreasing missing data (Lemmon et al., 2009; Wiens and Morrill, 2011), objective partitioning e.g. (Lanfear et al., 2012, Lanfear et al., 2014), model choice e.g. (Hoff et al., 2016) and application of the coalescent model (Kubatko and Degnan, 2007; McVay and Carstens, 2013). To fully exploit these options, the development of an independent data set to provide non-overlapping evidence should be undertaken (Campbell et al., 2014a). The Betancur et al. (2013a) dataset is a substantial resource that can provide a framework for the placement of lineages from broadly across the euteleost tree of life and integration with high-throughput sequence data (Campbell et al., 2017b; Sanciangco et al., 2016). However, it is also limiting as a base for further analysis, as the same underlying signal will be retained between studies, there are relatively few loci and a high degree of missing data is present. Our results show little agreement among analysis types regarding earlier-branching nodes and clear conclusions are difficult to make. An independent dataset of numerous loci to which the coalescent model and concatenated analysis can be applied will provide insight into pelagian interrelationships through concordance of the two analysis frameworks and concordance with results of previous studies e.g. (Campbell et al., 2017a). Similarly to early-branching euteleost relationships, a radiation of substantial age has occurred within Pelagia that may result in few informative characters to inform relationship being retained (Campbell et al., 2017a). Clear resolution of higher-level relationships within Pelagia may remain difficult even through the application of substantial effort; however, unclear parts of the fish Tree of Life have been clarified with the application of new datasets. Recent radiations of cichlids in East African rift lakes, for example, have long been problematic and only resolved clearly with new high-throughput sequencing generated datasets (Wagner et al., 2013). At a deeper time scale, flatfish monophyly within the larger carangimorph radiation was only weakly supported by mitogenomic data (Campbell et al., 2014b), and not supported or ambiguous with multilocus data sets that overlapped in terms of loci (Betancur et al., 2013a, 2013b; Betancur and Ortí, 2014; Campbell et al., 2013, 2014a). High-throughput sequence capture data demonstrated with enough effort, flatfish monophyly is robustly supported by DNA sequence data (Harrington et al., 2016). Therefore, generation of an independent and larger pelagian phylogenetic dataset may clarify many relationships within this radiation.

Acknowledgements

We would like to thank Thaddaeus Busser (Oregon State University) for reading our manuscript prior to submission and providing helpful comments. This study was supported by MEXT/JSPS KAKENHI grant numbers 22370035 and 26291083.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the

online version, at https://doi.org/10.1016/j.ympev.2018.03.008.

References

Allen, L.G., Cross, J.N., 2006. Surface waters. In: Allen, L.G., Pondella, D.J., Horn, M.H. (Eds.), The Ecology of Marine Fishes: California and Adjacent Waters. University of California Press, Berkely, pp. 320–341.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389.

Andrews, S., 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data.

Betancur, R.R., 2013. Data from: the tree of life and a new classification of bony fishes. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.c4d3j.

Betancur, R.R., Broughton, R.E., Wiley, E.O., Carpenter, K., López, J.A., Li, C., Holcroft, N.I., Arcila, D., Sanciangco, M., Cureton II, J.C., Zhang, F., Buser, T., Campbell, M.A., Ballesteros, J.A., Roa-Varon, A., Willis, S., Borden, W.C., Rowley, T., Reneau, P.C., Hough, D.J., Lu, G., Grande, T., Arratia, G., Ortí, G., 2013a. The tree of life and a new classification of bony fishes. PLoS Curr. http://dx.doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288.

Betancur, R.R., Li, C., Munroe, T.A., Ballesteros, J.A., Ortí, G., 2013b. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst. Biol. 62, 763–785. http://dx.doi.org/10.1093/sysbio/syt039.

Betancur-R, R., Wiley, E.O., Arratia, G., Acero, A., Bailly, N., Miya, M., et al., 2017. Phylogenetic classification of bony fishes. BMC Evol. Biol. 17 (1), 162. http://dx.doi.org/10.1186/s12862-017-0958-3.

Betancur, R.R., Ortí, G., 2014. Molecular evidence for the monophyly of flatfishes (Carangimorpharia: Pleuronectiformes). Mol. Phylogenet. Evol. 73, 18–22. http://dx.doi.org/10.1016/j.ympev.2014.01.006.

Betancur-R., Wiley, E.O., Bailly, N., Acero, A., Miya, M., Lecointre, G., Ortí, G., 2016. Phylogenetic Classification of Bony Fishes – Version 4.

Blair, C., Murphy, R.W., 2011. Recent trends in molecular phylogenetic analysis: where to next? J. Hered. 102, 130–138. http://dx.doi.org/10.1093/jhered/esq092.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btu170.

Brankovics, B., Zhang, H., van Diepeningen, A.D., van der Lee, T.A., Waalwijk, C., de Hoog, G.S., 2016. GRAbB: selective assembly of genomic regions, a new niche for genomic research. PLoS Comput. Biol. 12, e1004753. http://dx.doi.org/10.1371/journal.pcbi.1004753.

Campbell, M.A., Alfaro, M.E., Belasco, M., López, J.A., 2017a. Early-branching euteleost relationships: areas of congruence between concatenation and coalescent model inferences. PeerJ 5, e3548. http://dx.doi.org/10.7717/peerj.3548.

Campbell, M.A., Chen, W.-J., López, J.A., 2014a. Molecular data do not provide unambiguous support for the monophyly of flatfishes (Pleuronectiformes): a reply to Betancur-R and Ortí. Mol. Phylogenet. Evol. 75, 149–153. http://dx.doi.org/10.1016/j.ympev.2014.02.011.

Campbell, M.A., Chen, W.-J., López, J.A., 2013. Are flatfishes (Pleuronectiformes) monophyletic? Mol. Phylogenet. Evol. 69, 664–673. http://dx.doi.org/10.1016/j.ympev.2013.07.011.

Campbell, M.A., López, J.A., Satoh, T.P., Chen, W.-J., Miya, M., 2014b. Mitochondrial genomic investigation of flatfish monophyly. Gene 551, 176–182. http://dx.doi.org/10.1016/j.gene.2014.08.053.

Campbell, M.A., Nielsen, J.G., Sado, T., Shinzato, C., Kanda, M., Satoh, T.P., Miya, M., 2017b. Evolutionary affinities of the unfathomable Parabrotulidae: molecular data indicate placement of *Parabrotula* within the family Bythitidae, Ophidiiformes. Mol. Phylogenet. Evol. 109, 337–342. http://dx.doi.org/10.1016/j.ympev.2017.02.004.

Cavin, L., 2002. Effects of the Cretaceous-Tertiary boundary event on bony fishes. In: Geological and Biological Effects of Impact Events. Springer, pp. 141–158.

Chen, W.-J., Bonillo, C., Lecointre, G., 2003. Repeatability of clades as a criterion of reliability: a case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. Mol. Phylogenet. Evol. 26 (2), 262–288. http://dx.doi.org/10.1016/S1055-7903(02)00371-8.. ISSN 1055-7903, http://www.sciencedirect.com/science/article/pii/S1055790302003718.

Chen, W.-J., Mayden, R.L., 2010. A phylogenomic perspective on the new era of Ichthyology. BioScience 60, 421–432. http://dx.doi.org/10.1525/bio.2010.60.6.6.

Chevreux, B., Wetter, T., Suhai, S., 1999. Genome sequence assembly using trace signals and additional sequence information. In: German Conference on Bioinformatics. pp. 45–56.

Collette, B.B., Nauen, C.E., 1983. FAO species catalogue. Volume 2. Scombrids of the world. An annotated and illustrated catalogue of tunas, mackerels, bonitos and related species known to date. FAO Fisheries Synopses.

Doiuchi, R., Sato, T., Nakabo, T., 2004. Phylogenetic relationships of the stromateoid fishes (Perciformes). Ichthyol. Res. 51, 202–212.

Friedman, M., 2010. Explosive morphological diversification of spiny-finned teleost fishes in the aftermath of the end-Cretaceous extinction. In: Proc. R. Soc. Lond. B Biol. Sci. rspb20092177.

Friedman, M., 2009. Ecomorphological selectivity among marine teleost fishes during the end-Cretaceous extinction. Proc. Natl. Acad. Sci. 106, 5218–5223.

Friedman, M., Sallan, L.C., 2012. Five hundred million years of extinction and recovery: a Phanerozoic survey of large-scale diversity patterns in fishes. Palaeontology 55, 707–742.

Gladman, S., Seemann, T., 2012. VelvetOptimiser.

Haedrich, R.L., 1969. A new family of aberrant stromateoid fishes from the equatorial Indo-Pacific. Dana Rep 76, 1–14.

Harrington, R.C., Faircloth, B.C., Eytan, R.I., Smith, W.L., Near, T.J., Alfaro, M.E., Friedman, M., 2016. Phylogenomic analysis of carangimorph fishes reveals flatfish asymmetry arose in a blink of the evolutionary eye. BMC Evol. Biol. 16, 224. http://dx.doi.org/10.1186/s12862-016-0786-x.

Helfman, G., Collete, B., Facey, D., 1997. The Diversity of Fishes.

Hillis, D.M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47, 3–8.

Hoff, M., Orf, S., Riehm, B., Darriba, D., Stamatakis, A., 2016. Does the choice of nucleotide substitution models matter topologically? BMC Bioinf. 17, 143. http://dx.doi.org/10.1186/s12859-016-0985-x.

Horn, M.H., 1984. Stromateoidei: development and relationships. In: Richards, W.J., Cohen, D.M., Fahay, M.P., Kendall, A.W., Richardson, S.L. (Eds.), Ontogeny and Systematics of Fishes. Special Publication. American Society of Ichthyologists and Herpetologists, Lawrence, Kansas, pp. 620–628.

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T.P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., 2013. MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol. Biol. Evol. 30, 2531–2540.

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066. http://dx.doi.org/10.1093/nar/gkf436.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT multiple sequence alignment program. Brief. Bioinform. 9, 286–298. http://dx.doi.org/10.1093/bib/bbn013.

Knowles, L.L., Kubatko, L.S., 2011. Estimating Species Trees: Practical and Theoretical Aspects. John Wiley and Sons.

Konovalenko, I.I., Piotrovskiy, A.S., 1989. First description of a sexually mature Amarsipa, *Amarsipus carlsbergi*. J. Ichthyol. 28, 86–89.

Kubatko, L.S., Degnan, J.H., 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56, 17–24. http://dx.doi.org/10.1080/10635150601146041.

Lanfear, R., Calcott, B., Ho, S.Y.W., Guindon, S., 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol. 29, 1695–1701. http://dx.doi.org/10.1093/molbev/mss020.

Lanfear, R., Calcott, B., Kainer, D., Mayer, C., Stamatakis, A., 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14, 1–14. http://dx.doi.org/10.1186/1471-2148-14-82.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Meth. 9, 357–359. http://dx.doi.org/10.1038/nmeth.1923.

Lemmon, A.R., Brown, J.M., Stanger-Hall, K., Lemmon, E.M., 2009. The effect of ambiguous data on phylogenetic estimates obtained by Maximum Likelihood and Bayesian inference. Syst. Biol. 58, 130–145. http://dx.doi.org/10.1093/sysbio/syp017.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map (SAM) Format and SAMtools. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp352.

Maddison, W.P., Maddison, D.R., 2011. Mesquite: a modular system for evolutionary analysis.

Magnuson, J.J., 1979. Locomotion by scombrid fishes: Hydromechanics, morphology, and behavior. In: Hoar, W.S., Randall, D.J. (Eds.), Fish Physiology. Academic Press, London, pp. 239–313.

Magoč, T., Salzberg, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. http://dx.doi.org/10.1093/bioinformatics/btr507.

McVay, J.D., Carstens, B.C., 2013. Phylogenetic model choice: justifying a species tree of concatenation analysis. Phylogenetics Evol. Biol. 1. http://dx.doi.org/10.4172/2329-9002.1000114.

Mirarab, S., Reaz, R., Bayzid, M.S., Zimmermann, T., Swenson, M.S., Warnow, T., 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics 30. http://dx.doi.org/10.1093/bioinformatics/btu462.

Miya, M., Friedman, M., Satoh, T.P., Takeshima, H., Sado, T., Iwasaki, W., Yamanoue, Y., Nakatani, M., Mabuchi, K., Inoue, J.G., Poulsen, J.Y., Fukunaga, T., Sato, Y., Nishida, M., 2013. Evolutionary origin of the Scombridae (Tunas and Mackerels): members of a Paleogene adaptive radiation with 14 other pelagic fish families. PLoS ONE 8, e73535. http://dx.doi.org/10.1371/journal.pone.0073535.

Morgulis, A., Coulouris, G., Raytselis, Y., Madden, T.L., Agarwala, R., Schaffer, A., 2008. Database indexing for production MegaBLAST searches. Bioinformatics 24, 1757–1764. https://doi.org/0.1093/bioinformatics/btn322.

Nylander, J.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47–67. http://dx.doi.org/10.1080/10635150490264699.

Okamoto, M., Hoshino, K., Jintoku, T., 2011. First record of *Amarsipus carlsbergi* (Perciformes: Stromateoidei: Amarsipidae) from Japan and a northernmost range extension. Biogeography 13, 25–29.

Parin, N.V., Piotrovsky, A.S., 2004. Stromateoid fishes (suborder Stromateoidei) of the Indian Ocean (species composition, distribution, biology, and fisheries). J. Ichthyol. 44, S33.

Sanciangco, M.D., Carpenter, K.E., Betancur-R, R., 2016. Phylogenetic placement of enigmatic percomorph families (Teleostei: Percomorphaceae). Mol. Phylogenet. Evol. 94, 565–576. http://dx.doi.org/10.1016/j.ympev.2015.10.006.

Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33, 1654–1668. http://dx.doi.org/10.1093/molbev/msw079.

Shen, X.-X., Hittinger, C.T., Rokas, A., 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1 126 126 10.1038/

s41559-017-0126.
Slater, G.S.C., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinf. 6, 31. http://dx.doi.org/10.1186/1471-2105-6-31.
Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22, 2688–2690. http://dx.doi.org/10.1093/bioinformatics/btl446.
Stamatakis, A., Ott, M., 2008. Efficient computation of the phylogenetic likelihood function on multi-gene alignments and multi-core architectures. Philos. Trans. R. Soc. B Biol. Sci. 363, 3977–3984. http://dx.doi.org/10.1098/rstb.2008.0163.
Wagner, C.E., Keller, I., Wittwer, S., Selz, O.M., Mwaiko, S., Greuter, L., Sivasundar, A., Seehausen, O., 2013. Genome-wide RAD sequence data provide unprecedented resolution of species boundaries and relationships in the Lake Victoria cichlid adaptive radiation. Mol. Ecol. 22, 787–798. http://dx.doi.org/10.1111/mec.12023.
Wiens, J.J., Morrill, M.C., 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60, 719–731. http://dx.doi.org/10.1093/sysbio/syr025.
Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 18, 821–829. http://dx.doi.org/10.1101/gr.074492.107.
Zhang, C., Sayyari, E., Mirarab, S., 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. In: Meidanis, J., Nakhleh, L., (Eds.), Comparative Genomics. Springer International Publishing, pp. 53–75.
Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–598. http://dx.doi.org/10.1080/10635150290102339.