Confronting Sources of Systematic Error to Resolve Historically Contentious Relationships: A Case Study Using Gadiform Fishes (Teleostei, Paracanthopterygii, Gadiformes)

Adela Roa-Varón
Rebecca B. Dikow
(...)
Eric J. Hilton

Follow this and additional works at: https://scholarworks.wm.edu/vimsarticles

Part of the Genetics and Genomics Commons
Confronting Sources of Systematic Error to Resolve Historically Contentious Relationships: A Case Study Using Gadiform Fishes (Teleostei, Paracanthopterygii, Gadiformes)

ADELA ROA-VARÓN1, 2, 3, *, REBECCA B. DIKOW4, GIORGIO CARNEVALE5, LUKE TORNABENE6, 7, CAROLE C. BALDWIN2, CHENHONG LI8, AND ERIC J. HILTON3

1National Systematics Laboratory of the National Oceanic Atmospheric Administration Fisheries Service, 10th St. & Constitution Ave. NW, Washington, DC 20560, USA; 2Department ofVertebrate Zoology, National Museum of Natural History, Smithsonian Institution, 10th St. & Constitution Ave. NW, Washington, DC 20560, USA; 3Virginia Institute of Marine Science, William & Mary, Gloucester Point, VA 23062, USA; 4Data Science Lab, Office of the Chief Information Officer, Smithsonian Institution, 10th St. & Constitution Ave. NW, Washington, DC 20560, USA; 5Dipartimento di Scienze della Terra, Università degli Studi di Torino, Via Violetta Caluso, 35, 10125 Torino, Italy; 6School of Aquatic and Fishery Sciences, University of Washington, 1120 NE Boat Street, Seattle, WA 98105, USA; 7Burke Museum of Natural History and Culture, 4300 15th Ave NE, Seattle, WA 98105, USA; 8Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, 999 Hucheng Ring Rd, Pudong, Shanghai, China

*Correspondence to be sent to: National Museum of Natural History, Smithsonian Institution, Washington, DC 20560, USA, E-mail: Roa-VaronAk@si.edu.

Received 23 September 2019; reviews returned 27 November 2020; accepted 2 December 2020

Abstract.—Reliable estimation of phylogeny is central to avoid inaccuracy in downstream macroevolutionary inferences. However, limitations exist in the implementation of concatenated and summary coalescent approaches, and Bayesian and full coalescent inference methods may not yet be feasible for computation of phylogeny using complicated models and large data sets. Here, we explored methodological (e.g., optimality criteria, character sampling, model selection) and biological (e.g., heterochrony, branch length heterogeneity) sources of systematic error that can result in biased or incorrect parameter estimates when reconstructing phylogeny by using the gadiform fishes as a model clade. Gadiformes include some of the most economically important fishes in the world (e.g., Cod, Hakes, and Rattails). Despite many attempts, a robust higher-level phylogenetic framework was lacking due to limited character and taxonomic sampling, particularly from several species-poor families that have been recalcitrant to phylogenetic placement. We compiled the first phylogenetic data set, including 14,208 loci (>2.8 M bp) from 58 species representing all recognized gadiform families, to infer a time-calibrated phylogeny for the group. Data were generated with a gene-capture approach targeting coding DNA sequences from single-copy protein-coding genes. Species-tree and concatenated maximum-likelihood (ML) analyses resolved all family-level relationships within Gadiformes. While there were a few differences between topologies produced by the DNA and the amino acid data sets, most of the historically unresolved relationships among gadiform lineages were consistently well resolved with high support in our analyses regardless of the methodological and biological approaches used. However, at deeper levels, we observed inconsistency in branch support estimates between bootstrap and gene and site coefficient factors (gCF, sCF). Despite numerous short internodes, all relationships received unequivocal bootstrap support while gCF and sCF had very little support, reflecting hidden conflict across loci. Most of the gene-tree and species-tree discordance in our study is a result of short divergence times, and consequent lack of informative characters at deep levels, rather than incomplete lineage sorting. We use this phylogeny to establish a new higher-level classification of Gadiformes as a way of clarifying the evolutionary diversification of the order. We recognize 17 families in five suborders: Bregmacerotoidei, Gadoidae, Ranicipitoidae, Merluccioidei, and Macrouroidei (including two subclades). A time-calibrated analysis using 15 fossil taxa suggests that Gadiformes evolved ~79.5 Ma in the late Cretaceous, but that most extant lineages diverged after the Cretaceous–Paleogene (K-Pg) mass extinction (66 Ma). Our results reiterate the importance of examining phylogenomic analyses for evidence of systematic error that can emerge as a result of unsuitable modeling of biological factors and/or methodological issues, even when data sets are large and yield high support for phylogenetic relationships. [Branch length heterogeneity; Codfishes; commercial fish species; Cretaceous–Paleogene (K-Pg); heterochrony; systematic error; target enrichment.]
Gadus chalcogramma

Alaskan Pollock (on the status of some stocks (FAO 2016). For example, climate change among others, and worrisome reports habitat modification, ocean acidification, nanoparticles, recreational fisheries face from overfishing, physical important considering the threats that commercial and state of their higher systematics and the large number of

Hake (fully to overfished in the northeast Atlantic, and all

morhua

Gadus fully fished in the North Pacific, Atlantic Cod (catch of marine fishes (FAO 2016). However, the confused commercially harvested fishes in the world (e.g., Alaskan fishes as a model clade.

confronting sources of systematic error using gadiform biases (e.g., Davalos et al. 2008). Applications of these advances in genome-scale data set production and phylogenetic inference offer an opportunity to improve our knowledge of the systematic relationships of organisms. In this study, we explore approaches for confronting sources of systematic error using gadiform fishes as a model clade.

Gadiformes include some of the most important commercially harvested fishes in the world (e.g., Alaskan Pollock, Atlantic and Pacific Cods, Blue Whiting, and Hake), accounting for more than a fifth of the world’s catch of marine fishes (FAO 2016). However, the confused state of their higher systematics and the large number of poorly known species obscure the necessary framework for many conservation initiatives. This is particularly important considering the threats that commercial and recreational fisheries face from overfishing, physical habitat modification, ocean acidification, nanoparticles, climate change among others, and worrisome reports on the status of some stocks (FAO 2016). For example, Alaskan Pollock (Gadus chalcogramma) is considered fully fished in the North Pacific, Atlantic Cod (Gadus morhua) is overfished in the Northwest Atlantic and fully to overfished in the northeast Atlantic, and all Hake (Merluccius spp.) stocks are considered overfished.

Gadiformes inhabit cool waters circumglobally and are found in all portions of the water column at high latitudes but are primarily restricted to deeper layers of tropical seas. Between 11 and 14 families, 84 genera, and 613 species are currently recognized (e.g., Endo 2002; Roa-Varón et al. 2009; Nelson et al. 2016). The order has been characterized as monophyletic, even though well-defined morphological synapomorphies supporting its monophyly have yet to be established (e.g., Rosen and Patterson 1989; Patterson and Rosen 1989; Murray and Wilson 1999; Endo 2002).

Gadiformes appear to be nested within the Paracanthopterygii; however, the taxonomic content of the superorder is the subject of debate based on both morphological (e.g., Greenwood et al. 1966, Patterson and Rosen 1989, Johnson and Patterson 1993, Davesne et al. 2016) and molecular (e.g., Miya et al. 2001, 2003, 2005, 2007; Grande et al. 2013; Near et al. 2013; Betancur-R et al. 2013; Chen et al. 2014; Alfaro et al. 2018; Hughes et al. 2018) data. There is, however, a general agreement among recent molecular studies for a lineage comprising Zeiformes, Stylephoriformes, and Gadiformes. Among these orders, Gadiformes are the most species rich and phylogenetically complex. The resolution of the interrelationships of Gadiformes, therefore, is critical for better understanding the relationships and evolution of paracanthopterygians and other early divergent acanthomorph fishes.

There have been several attempts to resolve the relationships of Gadiformes based primarily on morphological data (Berg 1940; Svetovidov 1948; Fahay and Markle 1989; Howes 1989; Markle 1989; Nolf and Steurbaut 1989) (S1a-c, Supplementary Material available on Dryad). These studies produced largely conflicting results, with no consensus regarding the phylogeny and classification of the group (Cohen 1989). Endo (2002) published the most recent and extensive morphology-based analysis of the phylogeny and classification of Gadiformes (S1d, Supplementary Material available on Dryad). Several genetic studies have focused on the systematic relationships of subgroups (i.e., families or the phylogenetic affinities of specific taxa) of Gadiformes (e.g., Paulin 1983; Dunn and Matase 1984; Carr et al. 1999; Roldan et al. 1999; Quinteiro et al. 2000; Grant and Leslie 2001; Møller et al. 2002; Torii et al. 2003; Teletchea et al. 2006). To date, Roa-Varón et al. (2009) published the only molecular phylogeny focused on Gadiformes based on one nuclear and two mitochondrial DNA (mtDNA) genes from 117 taxa (S1e, Supplementary Material available on Dryad). Other studies, while not directly addressing the phylogeny of Gadiformes, presented results based on substantial yet incomplete taxon sampling. In an analysis of all bony fishes, Betancur-R et al. (2013) used 21 molecular markers for 42 taxa representing nine families of Gadiformes (S1f, Supplementary Material available on Dryad).
Malstrøm et al. (2016) analyzed 567 exons representing 111 genes (71,418 bp) for 27 gadiform species from seven families. The authors reported that Gadiformes lost the major histocompatibility complex II approximately 105 Ma (S1g, Supplementary Material available on Dryad). Hughes et al. (2018) mined data from Malstrøm et al. (2016) to compile 1,105 protein-coding genes for 11 gadiform families (S1h, Supplementary Materials available on Dryad), but no consensus regarding the relationships among these families emerged from a variety of analyses (Hughes et al. 2018, Supplementary Figs. S2–S5). These prior studies have yielded little to no consensus on the family level relationships of Gadiformes. Here, we generated new genomic data for all families of the order using a custom-designed probe set targeting more than 14,000 loci in the nuclear genome. We inferred the species tree based on concatenated and multispecies coalescent methods for data sets differing in the amount of missing data and loci informativeness to investigate biological (e.g., compositional heterogeneity, heterotachy, branch length heterogeneity) and methodological (e.g., optimality criteria, character sampling, model selection) sources of systematic error to infer the most robust and comprehensive backbone phylogeny for Gadiformes to date, estimate divergence times, and highlight remaining challenges.

**Materials and Methods**

### Species Sampling and Molecular Techniques

We collected phylogenomic data from single-copy nuclear coding sequence markers across 58 species, including 51 gadiforms representing all currently recognized families and subfamilies (S2a, Supplementary materials available on Dryad), using a targeted sequencing approach and the bioinformatic workflows of Yuan et al. (2016), Song et al. (2017), and Jiang et al. (2019) with minor modifications (lab protocol, bioinformatics pipeline, scripts, and associated files available from the Dryad Digital Repository [https://doi.org/10.5061/dryad.5qfzd2z]). Because the relationships among the earliest diverging acanthomorphs are unresolved (Near et al. 2013; Betancur-R et al. 2013; Chen et al. 2014; Alfaro et al. 2018), we included representatives of all putative acanthomorph orders (Lampridiformes, Percopsiformes, Polynemiformes, Stylephoriformes, and Zeiformes) and Ophidiiformes (as representative of all remaining percomorphs) as outgroups. We generated genomic data for 43 species. For the remaining 15 species, loci of interest were extracted from published genomic data (Malstrøm et al. 2016) following a modified pipeline for harvesting loci from genomes (Faircloth 2016). Lab protocols and bioinformatics are described briefly below and detailed in S3–S5, Supplementary Materials available on Dryad.

### Gene Capture and Probe Design

An initial set of 17,817 conservative single-copy nuclear coding sequence regions (CDS > 90 bp) identified by Song et al. (2017) were obtained by comparing eight well-annotated fish genomes (Anguilla japonica, Danio rerio, Gadus morhua, Gasterosteus aculeatus, Lepisosteus osseus, Oreochromis niloticus, Orezyxias latipes, and Tetrazyon nigrovirilis) using Evolmarkers (Li et al. 2013). Common metrics to assess assembly quality are scaffold and contig N50 and L50, which indicate the total number and minimum length, respectively, of all scaffolds or contigs that together account for 50% of the genome. The genomes were highly contiguous: contigs with N50 range: 11 Kb to 2.9 Mbp and L50 range 88–19 kb; scaffold with N50 range: 734 Kb–38.8 Mbp and L50 range 1–102 (S3, Table 1 Supplementary Materials available on Dryad). A set including 14,217 CDS (>120 bp) was generated and used to design baits based on the sequences of G. morhua (no degenerated baits were included). Bait sequences of 120 bp were tiled to obtain ≥2× coverage of each targeted locus (60 bp overlap between baits). Biotinylated RNA probes of bait sequences were synthesized by Arbor Bioscience (formerly MYcroarray, Ann Arbor, MI, USA). Illumina sequencing libraries (Meyer and Kircher 2010) were prepared for each sample following the “with-bead” method of Li et al. (2013).

### Data Assembly, Orthology Testing, and Alignment

Data were assembled using a pipeline modified from Li et al. (2013, S5, Supplementary Materials available on Dryad). We removed adapter sequences, low-quality reads, and PCR duplicates, and parsed reads into files based on similarity to the target loci. Paired reads were assembled into contigs and putatively orthologous genes were identified by matching contigs to the G. morhua genome using BLAST + (v. 2.4.0; Camacho et al. 2009). Loci were translated into amino acids (AA), and both DNA and AA loci were aligned in MAFFT (v. 7.221; Katoh and Standley 2013) and concatenated with FASconCAT-G (Kück and Longo 2014).

### Concatenated Phylogenetic Analyses

In this study, we used numerous approaches to tackle potential sources of systematic error from inappropriate modeling of biological and/or methodological factors (Fig. 1). Both of these sources of error can lead to incorrect parameter estimates in phylogenomic analyses yet result in statistically highly supported, phylogenetic trees.

**Missing Data and Loci Informativeness.—**We built four matrices to account for missing data and loci informativeness. The first matrix is the original data set (DNA sequences, 58 taxa, hereafter as DNA-S8T). The remaining three matrices were built with MARE v0.1.2-rc
**FIGURE 1.** Workflow of the data analyses exploring biological (solid) and methodological (dotted) sources of systematic error that can result in biased or incorrect parameter estimates in our phylogenomic analyses. *degen applied.*

(MAtrix REduction, Meyer et al. 2011) by adjusting the α value (α1 more relaxed and α5 stricter), 58T-SOS-1: α = 1; 58T-SOS-3: α = 3; and 58T-SOS-5: α = 5 to search for the highest information content while excluding as few loci as possible (Fig. 1). MARE combines potential signal (i.e., informativeness) of loci with its data coverage to generate Selected Optimal Subsets (SOS) of taxa and loci. The relative information content of each locus was calculated as the average value over all taxa including missing taxa, whereas the total average information content of the supermatrix was calculated as the sum of the relative information content of all loci in relation to the number of taxa. ML analyses were conducted using RAxML-NG v0.8.1 BETA (Kozlov 2018) under the GTR + G model. ML tree searches were conducted using 20 distinct random starting trees (10 random + 10 parsimony). Support for nodes in the ML tree was assessed with 100 slow bootstrap replicates for each of the four matrices. We define the support of the species trees as follows: strongly supported nodes have a BS > 90, moderately supported nodes are between 90 and 70 BS, and poorly supported nodes are those with a BS < 70. All trees were visualized with FigTree v1.4.2 (Rambaut 2009).

Data Partitioning and Model Selection.—The following two strategies were applied to the optimized data sets (58T-SOS-1) to evaluate the impact of data partitioning and model selection: (i) unpartitioned versus partitioned by codon position with the GTR + G model; ML tree searches were conducted in RAxML-NG v0.8.1 using the same parameters described above; and (ii) partitioning by codon position was used as input for model selection in IQ-TREE v1.6 (Kalyaanamoorthy et al. 2017). The best-fitting model for each codon position was selected based upon Akaike Information Criteria (Akaike 1974).

Branch Length and Compositional Heterogeneity.—Outgroup taxa may attract fast-evolving species of the ingroup through long branch attraction (LBA), thereby forcing the rapidly evolving taxa to be recovered as emerging too deeply in the tree (Philippe et al. 2011; Whelan et al. 2015). This effect was assessed by: (i) removing distant outgroups (Lampridiformes and the Ophidiiformes; DNA-54T) followed by applying the same matrix reduction strategy applied to 58T data set with MARE v0.1.2-rc (MATRix REDuction, Meyer et al. 2011); (ii) removing all outgroups (DNA-51T); (iii) building a reduced matrix including only the 146 loci found in Bregmaceros spp. (DNA-58T-R); and (iv) removing long-branch taxa (Bregmaceros spp. DNA-56T). ML tree searches were conducted in RAxML-NG v0.8.1 using the same settings as described above (Fig. 1).

Heterotachy.—Because heterotachy is one of the primary sources of model misspecification (Kolaczkowski and Thornton 2004; Crotty et al. 2017), we addressed the rate of variation across sites and lineages using the following approaches: (i) edge-proportional partition model with proportional branch lengths (edge option);
partition has its own partition specific rate (-spp option); (ii) edge-unlinked partition model—each partition has its own set of branch lengths (-sp option); and (iii) General Heterogeneous evolution On a Single Topology (GHOST) model using GTR+FO+I14 in IQ-TREE (Nguyen et al. 2015; Crotty et al. 2019). The first two approaches were partitioned a priori by codon position, and then separate branch length and/or model parameters were inferred for each partition. However, because the partition itself is a possible source of model misspecification, the GHOST model offers the advantage of inferring heterotachous evolutionary processes without specifying partitions a priori and minimizing model assumptions (Fig. 1).

Saturation and Base Composition Bias.—Rapidly evolving synonymous nucleotide changes are mostly uninformative at deep levels and often cause analytical problems such as LBA and nucleotide compositional heterogeneity, among others (e.g., Song et al. 2010, Zwick et al. 2012). To examine the extent of saturation for the nucleotide data sets (DNA-54T, SOS-1) the degen v1.4 Perl script (Zwick et al. 2012) was used to exclude synonymous signal that can contribute to saturation. ML tree searches were conducted in RAxML-NG v0.8.1 under the GTR+G model and using the same settings as described above (Fig. 1). Because translation to amino acids can eliminate potentially problematic synonymous nucleotide changes a tree based on amino acids for AA-58T data set was inferred by estimating the number of partitions with PF2 (Lanfear 2016) and the following parameters: “rcluster-max” to 1000 and “rcluster-percent” to 10 for the relaxed clustering algorithm. Model selection was performed in IQ-TREE followed by ML analyses using RAxML-NG v0.8.1 following settings as above (Fig. 1). Because translation to amino acids can eliminate potentially problematic synonymous nucleotide changes a tree based on amino acids for AA-58T data set was inferred by estimating the number of partitions with PF2 (Lanfear 2016) and the following parameters: “rcluster-max” to 1000 and “rcluster-percent” to 10 for the relaxed clustering algorithm. Model selection was performed in IQ-TREE followed by ML analyses using RAxML-NG v0.8.1 following settings as above (Fig. 1).

Coalescent Species Tree Analyses

Gene Tree Heterogeneity.—Species trees were inferred for the DNA-58T, 58T-SOS-1, and the AA-58T data sets using the coalescent summary method in ASTRAL-II v5.6.3 (Mirarab and Warnow 2015). The individual input gene trees were first estimated by using RAxML-NG v0.8.1 as described above. We used the best-fitting ML gene trees as an input and conducted gene/site resampling on the bootstrap replicates (Fig. 1). We collapsed to polytomies with very low bootstrap support (< 1 BS, <10 BS, <20 BS, <30 BS) within each gene tree to discern if discordant topologies are due to low informativeness of the data and to minimize the impact of estimation error (Zhang et al. 2017, 2018; Sayyari and Mirarab 2018).

Gene Tree and Species Tree Support
To assess the degree of discordance between gene trees and the species tree in our data, we applied the following approaches. First, because bootstrap or posterior probabilities do not provide a comprehensive measure of the underlying agreement or disagreement among sites and genes for any topology, we estimated the gene concordance factor (gCF) and the site concordance factor (sCF) for each 58T-SOS data set using IQ-TREE v1.6.10 (Kalyaanamoorthy et al. 2017). For every branch of the tree, gCF is the percentage of “decisive” gene trees containing that branch and accounting for variable taxon coverage among gene trees and sCF is the percentage of “decisive” sites supporting a branch in the tree. sCF was calculated based on the concatenated alignment and gCF based on the individual gene trees (Fig. 1). Both gCF and sCF complement bootstrap branch support values by providing a full description of underlying disagreement among loci and sites (Minh et al. 2020). Bootstrap, gCF, and sCF statistics for each branch were plotted in R (R Core Team 2013). Third, to test if the discordance among gene trees and sites come from ILS and/or poorly estimated gene trees, we calculated the probability that the data can reject equal frequencies for genes (gEFp) and for sites (sEFp) using a chi-square test (assuming that sites are not linked within single loci) to see if they are significantly different. If they were not significant, discordance among gene trees and among sites was likely due to ILS (script available here: http://www.robertlanfear.com/blog/files/concordance_factors.html). Second, to test the significance of the differences between phylogenies derived from our analyses, we employed the approximately unbiased (AU) test assuming independence of the sites (Shimodaira 2002) included in IQ-TREE v1.6.10 (Kalyaanamoorthy et al. 2017)). We may reject the possibility that a tree is the most likely tree among all candidates when the AU p-value <0.05.

Divergence Time Analyses

Divergence-time estimation was conducted using the program RelTime (Tamura et al. 2012), a fast, ad hoc approach for large molecular phylogenies implemented in MEGA v. 7.0.20 (Kumar et al. 2012, 2018). The basis of the RelTime approach is a relative rate framework that combines comparisons of evolutionary rates in sister lineages with the principle of minimum rate change between evolutionary lineages and their respective descendants (Sanderson 1997; Tamura et al. 2012, 2018). This approach relaxes the molecular-clock assumption and smooths the local rates in a tree (Kumar and Hedges 2016b). Other ML-based approaches (e.g., TreePL) are computationally efficient, requiring only a fraction of the time and resources demanded by Bayesian approaches and generate accurate time estimates (Tamura et al. 2012, 2018). However, they are limited by the lack of reliable calculation of the uncertainty surrounding divergence times, which are represented by confidence intervals (CIs) implemented in RelTime (Tao et al. 2018). In total, we used 15 calibration points across the best ML tree hypothesis using minimum and/or maximum boundaries under the GTR+G model and using all sites.
Justification and description of the fossils used for the calibration can be found in S6, Supplementary Material available on Dryad.

RESULTS AND DISCUSSION

Gene Capture Data Collection

We analyzed sequences for 14,208 loci from 58 taxa: 51 species representing all families and subfamilies of Gadiformes and seven outgroups representing all earliest diverging acontomorph orders and other percomorphs. On average, 36.5% of loci were recovered (95% CI 3.5–60.8%), locus length was 193.7 bp (95% CI 167.4–216.5%), there were 285,599 parsimony-informative sites (95% CI 18,557–390,582), and there were 31.4% variable sites (95% CI 25.7–38.4%). The number of sequences generated, accession numbers, number of raw reads, contigs and values to assess the quality of the assembled loci, can be found in S2a, Supplementary Materials available on Dryad.

The percentage of capture efficiency by family ranged from 12.3% to 76.0%, except for Bregmacerotidae, which registered the lowest capture efficiency of 4.8%. The composition and properties of the primary matrices are summarized in S2b, Supplementary Materials available on Dryad. Data available from NCBI (http://www.ncbi.nlm.nih.gov/bioproject/609091) and TreeBASE (http://purl.org/phylo/treebase/phylows/study/TB2:S27320).

Phylogenetic Analyses

Phylogenetic hypotheses were inferred using ML and coalescent summary methods based on 12 different data sets to assess how methodological and biological sources of systematic error may affect phylogenetic inference (S2b, Supplementary Materials available on Dryad).

Missing Data and Loci Informativeness.—Four distinct supermatrices were assembled. The complete unreduced matrix (DNA-58T) was based on all filtered alignments comprising at least four sequences. Three “MARE” matrices were obtained by removing relatively uninformative and low-coverage loci from the DNA-58T matrix by adjusting the alpha variable: α1: 58T-SOS-1 (8,244 loci; 1,784,739 sites), α2: 58T-SOS-3 (4,933 loci; 1,052,865 sites), and α5: 58T-SOS-5 (3,551 loci; 715,188 sites), while keeping the number of taxa fixed (deleting organisms was allowed, but none was excluded; S7, Supplementary Materials available on Dryad). The ML phylogenetic trees were inferred using a single GTR+G model across the alignment for all four matrices (Fig. 2a–d).

A fully congruent topology was resolved for DNA-58T and 58T-SOS-1 (97% and 99% of the nodes with more than 95% BS values respectively; Fig. 2a-b). Relationships among the earliest diverging taxa are highly supported, as are all clades corresponding to families and subfamilies. The tree was rooted on Ophidiiformes, and Lampridiformes were recovered as the sister group to (Polygymniformes, (Perciformes, (Zoiformes, (Oestinephalliiformes, Gadiformes)))). Among the taxa sampled Stylephorus chordatus was recovered as the closest extant relative of Gadiformes. Within Gadiformes, six main lineages were found based on the concatenated ML topology (2.5–24.4 gCF, 33.2–41.6 sCF, and 99–100% BS) representing 17 families (45.6–83.0 gCF; except Lotidae with 11.2; 51.4–89.6 sCF; and 100% BS). Bregmacerotidae was resolved on a long branch as the sister group of all other gadiforms, which were divided into two well-supported clades (Fig. 2a–d). The first clade, Gadoidei, includes (Phycidae, (Gaidropsaridae, (Lotidae, Gadidae))). Within the second clade, the monotypic Ranicipitidae and Merlucciidae were successive sister clades to the remaining Gadiformes, which are represented here by Macrouroids with two subclades. The first of these subclades comprises (Trachyrinoridae, (Euclichthysiidae, (Melanomidae, Muraenolepididae))) and the second (Bathygadidae, (Macruronidae, Lyconidae), (Steindachneriidae, Macrouridae), with Moridae as its sister group. Both ML topologies (58T-SOS-3 and 58T-SOS-5) were largely consistent with those obtained from the ML analysis of the DNA-58T and 58T-SOS-1 and differed only in one clade in the second lineage that registered the lowest gCF, sCF, and bootstrap support values: (Muraenolepididae, Melanomidae) was found to be sister to the remaining gadiforms in 58T-SOS-3 (2.3/35.1/53, respectively; Fig. 2c; S2d, Supplementary Materials available on Dryad) or nested within a clade including (Euclichthysiidae, (Trachyrinoridae, (Melanomidae, (Muraenolepididae, Merlucciidae)))) in the 58T-SOS-5 matrix (0.3/37.2/21, respectively; Fig. 2d; S2d, Supplementary Materials available on Dryad). The drop in the concordance factors and bootstrap support values in the 58T-SOS-3 and 58T-SOS-5 matrices, mainly in the second lineage are possibly due to the decrease of information content. This corroborates the importance of exploring different α values when running MARE to identify the matrix with the highest information content while excluding as few loci as possible. The preferred topology among the four matrices in terms of missing data and loci informativeness is the topology resulting from the 58T-SOS-1. The matrix is about 58% of the size of the original, the information content increased from 0.27% to 40%, and was supported by the combination of concordance factors and bootstrap support values (S2d, Supplementary Materials available on Dryad).

Selected Removal of Outgroup and Long-Branch Taxa.—A topology that is fully congruent with DNA-58T was found after: (i) removing the most distant outgroup taxa (DNA-54T; Fig. 3a,b); (ii) applying a similar strategy with MARE (54T-SOS-1; S8, Supplementary Materials available on Dryad); (iii) removing the long branches (two Bregmaceros spp.-DNA-56T; Fig. 3d); and (iv) removing all the outgroups (DNA-51T; Fig. 3c). Nodes with less than 95% BS increased from 2.8% to 3.7%
and 4.7% in the last two approaches, respectively. Both increase and reduction in branch support have been reported after removal of taxa, suggesting other sources for systematic error such as improper modeling of biological phenomena (e.g., compositional bias, rate heterogeneity, heterotachy) and/or methodological issues (e.g., poor model selection, biased taxon or gene sampling, poor orthology inference) (e.g., Philippe et al. 2011; Whelan et al. 2015). We attempted to break up the long branches by increasing taxon sampling within Bregmacerotidae through the inclusion of loci from *B. cantori* that were harvested from Malstrøm et al. (2016, Supplementary Figure 1; a long branch leading to Bregmacerotidae was also observed in their study). However, very few loci were captured for *B. cantori* (0.2% capture efficiency representing only 25 of our 14,208 loci). We did not include the species in the analyses because it would introduce more noise than phylogenetic signal and increase computing time due to the amount of missing data.

The topology recovered through analysis of the reduced data set, which included only those loci present in Bregmacerotidae (146 loci recovered; 83,025 bp; 18,557 parsimony informative sites), changed the topology drastically (S9, Supplementary Materials available on Dryad). For example, with the reduced data set, Melanonidae, Moridae, Euclichthyidae, and Trachyrincidae form a series of successive sister taxa to all the remaining gadiforms, and Merluccidae and Ranicipitidae are nested within the Gadoidei. Additionally, bootstrap support dropped at many nodes, suggesting significant loss of phylogenetic signal. This is not unexpected because the most consistently recovered loci across taxa are those that are more conserved and therefore less informative.

**Codon-based Partitioning, Substitution and Heterotachy Models.**—As a result of the optimization process, the 5S-T-SOS-1 and 54S-T-SOS-1 matrices were used in the downstream phylogenetic analyses to evaluate the impact of data partitioning and model selection. For each of the following strategies, we found a fully congruent topology: (a) unpartitioned with GTR+G model; (b) partitioning scheme by codon position with the GTR+G model; and (c) partitioning scheme with substitution
FIGURE 3. Branch length and compositional heterogeneity. Maximum likelihood inference using concatenated data sets: a) DNA-58T (14,208 loci/2,874,858 bp); b) DNA-54T-SOS-1 (removing the most distant outgroups); c) DNA-51T (all outgroups removed); d) DNA-56T (matrix where the long-branch taxa were removed) Bootstrap support values are color coded.
model and rate of evolution by codon position estimated with IQ-TREE v1.6 (P1: GTR+F+R4; P2: GTR+F+R5; P3: GTR+F+R7) (S10, Supplementary Materials available on Dryad). Our results suggest that partitioning by codon position and choosing the best-fitting model by codon position had little impact on the resulting tree topology. The models estimated by IQ-TREE for each codon position were the most parameter rich and very similar to GTR+G (58T-SOS-1—P1: GTR+F+R4; P2: GTR+F+R5; P3: GTR+F+R7 and 54T-SOS-1—P1: GTR+F+R4; P2 and P3: GTR+F+R6)). This could result in consistent topologies with marginal differences in bootstrap support values.

Additional analyses to accommodate heterotachy resulted in the same topology and LBA artifact for the 58T-SOS-1 independent of the approach used (edge-proportional, edge-unlinked, and GHOST models) (S11a-c, Supplementary Materials available on Dryad). In contrast, the phylogenetic placement of three families changed depending on the model used for the 54T-SOS-1 matrix (S11d–f, Supplementary Materials available on Dryad). The edge-unlinked partition model suggested Muraenolepididae + Merlucciidae (sister to the rest of Gadiformes except Bregmacerotidae), Trachyrincidae within Macrouroidei—clade 2, and Ranicipitidae within Gadoidi; edge-proportional partition model recovered Muraenolepididae (sister to the rest of Gadiformes except Bregmacerotidae); and the GHOST model, included Ranicipitidae within Gadoidi (S11d–f, Supplementary Materials available on Dryad). The different topologies between the data sets suggest that the number of taxa included in the outgroup is affecting the performance of the models addressing the rate of variation across sites and lineages and/or that those models of molecular evolution are insufficient to deal with all the phenomena present in these genome-scale matrices.

Saturation and Base Composition Bias.—The topology resulting from the DNA-54T + degenerate amino acids data set (S12, Supplementary Materials available on Dryad). Analyses using the concatenated amino acid alignment for addressing problematic synonymous nucleotide changes consisted of 956,264 sites and 107 subsets resulting from Partition Finder 2 (PF2, Lanfear 2016) using the JTT+F+R5 model. ML analysis generated phylogenetic hypotheses in which the majority of nodes were well supported, with some exceptions (S13, Supplementary Materials available on Dryad). We observed topological conflict between the amino acid (AA-58T) and DNA ML trees (58T-SOS-1) at only three nodes: (i) the placement of Merlucciidae as the sister group of all gadiforms except Bregmacerotidae with high support (99% BS); (ii) Ranicipitidae within the gadoidei clade with moderately high support (90% BS); and (iii) Euclischthyidae sister to ((Melanomidae, Muraenolepididae), Trachyrincidae) with relatively low support (77% BS) in the AA-58T data set. Recovering Bregmacerotidae, Merlucciidae and Muraenolepididae diverging so deeply in the amino acid-based topology suggests low amino acid variation and/or that LBA could be biasing the topology. However, using amino acids instead of nucleotides has been suggested to alleviate LBA (Inagaki et al. 2004; Mathews et al. 2010; Talavera and Vila 2011). Therefore, we favor the explanation of low phylogenetic signal in the amino acid data set.

Gene Tree Heterogeneity.—The species trees inferred using ASTRAL (DNA-58T and 58T-SOS-1) resulted in the same topology and recovered monophyly of all gadiform families (except Lotidae) as in the concatenated analyses DNA-58T, 58T-SOS-1, DNA-54T, and the 54T-SOS-1 with high support (S14, Supplementary Materials available on Dryad). The nonmonophyly of Lotidae had low support compared with the ML tree estimated for the 58T-SOS-1 (gCF = 11.16%; sCF = 45.67% and 100% BS). After collapsing branches with low support (<1% and >10% support) within each gene tree the monophyly of Lotidae still was not supported. However, when branches with >20% and <30% support were collapsed the monophyly of Lotidae was recovered with high support suggesting an increase in accuracy by reducing noise (S15, Supplementary Materials available on Dryad). In these trees, the families Merlucciidae and Muraenolepididae were recovered sister to the Gadoidi and Macrouridi lineages, as in the topologies recovered in the amino acid-based (S13, Supplementary Materials available on Dryad), and applying the edge-unlinked partition model from the 54T-SOS-1 data set (S11, Supplementary Materials available on Dryad). The topology resulting from the amino acids data set (AA-58T) included one clade of species, representing different families, that correspond to the sequences harvested from Malstrom et al. (2016), suggesting issues with the reading frame of these data (S16, Supplementary Materials available on Dryad). Additionally, it is likely that the individual amino acid trees contain insufficient phylogenetic information as a result of their short length.

Gene Tree and Species Tree Support
Concordance factors.—Our estimates of sCF and gCF were correlated across the species tree of gadiforms, but we noted that both of these measures fell well below those of bootstrap support measures (S2d, S17; Supplementary Materials available on Dryad). The topology resulting from the amino acids data set (AA-58T) included one clade of species, representing different families, but discordance at deeper nodes suggesting the monophyly of all families, but discordance at deeper nodes (e.g., suborders) increase. Short internodes connecting long branches at deep levels are suggestive of ancient divergences.
that gave rise to speciation events over relatively short periods of time (S2d, S17; Supplementary Materials available on Dryad).

The phylogenetic placement of Bregmacerotidae did not change in any of our analyses regardless of the methodological and biological approaches used to break the long branch (Figs. 2–5, S9–S16). For example, if this branch was being artificially attracted toward outgroups, a different placement for the family would be expected when different outgroups are included. Applying models of sequence evolution accounting for base compositional heterogeneity and heterotachy in partitioned and unpartitioned data sets recovered identical relationships, indicating that model choice did not bias results (S10, Supplementary Materials available on Dryad). Overall, findings presented here support Bregmacerotidae as sister to all remaining gadiform fishes (23.5 gCF; 49.9 sCF; 100 BS) and provide strong evidence that LBA artifacts are not biasing the placement of the family. However, support values can be explained by both good support and alternatively due to systematic error consistently affecting all loci. Additionally, the low capture efficiency (4.6%) for the family can exacerbate systematic error leading to LBA artifacts by reducing the ability to detect multiple substitutions, along these long branches. This limited our conclusions and future investigations including additional taxa will be needed to more fully assess its placement.

The historically unresolved relationships among gadiform lineages were consistently well resolved in our analyses regardless of the methodological and biological approaches used. Only three analyses rendered different placements for Merlucciidae, Muraenolepididae, and Rancipitidae: (i) coverage and loci informativeness analysis, in which Merlucciidae and Muraenolepididae were sister taxa in 58T-SOS-3 and 58T-SOS-5, but this hypothesis was poorly supported (5.5/2.9 gCF; 34.3/34.2 sCF; 59/30 BS, respectively; Fig. 2c,d; S2d, Supplementary Materials available on Dryad); (ii) amino acid-based analysis, in which Merlucciidae was recovered as the sister group of all gadiforms except Bregmacerotidae (S13, Supplementary Material available on Dryad); and (iii) heterotachy analysis in 54T-SOS-1, in which Muraenolepididae and Muraenidae were recovered sister to the Gadoidei and Macrourouidei lineages (edge-proportional partition); Muraenolepididae sister to all gadiforms except Bregmacerotidae (edge-unlinked partition model); and Rancipitidae was the earliest branch within Gadoidei according to GHOST (S11f, Supplementary Materials available on Dryad). The first two could be a result of loss of phylogenetically informative characters, while the third one suggests that theoretical work is needed to explore the relationship between outgroup rooting and LBA, especially with regard to how different outgroup taxon sampling strategies affect the probability of LBA and other systematic errors. Finally, most of the gene-to-species-tree discordance were not described by ILS, where 75.0% of nodes show significant differences between the two discordant topologies (S2e, Supplementary Material available on Dryad). Similarly, sCF shows 36.0% of nodes have significance difference among discordance sites. Hence, most of the gene-tree and species-tree discordance in our data set is likely due to short times between divergences and consequent lack of informative characters at deep levels rather than as a result of ILS.

**Gene Tree and Species Tree Incongruence**

Approximately unbiased test. —AU p-values represent how strongly competing topological hypotheses conflict with a preferred hypothesis, with a canonical value of 0.05 reflecting strong support for rejection. Alternative hypotheses to the most frequent best-fit tree were all rejected by AU test and therefore the phylogeny presented in Fig. 4 remains un falsified (S18; Supplementary Materials available on Dryad).

**Phylogeny and Evolution of Gadiformes (Figs. 4 and 5)**

Recent molecular phylogenetic studies have produced conflicting hypotheses of relationships among early diverging acanthomorphs (e.g., Lampridiformes, Polynemiformes, Percopsiformes, Zeiiformes, Stylephoriformes, and Gadiformes). While resolving the relationships among acanthomorphs is beyond the scope of this study, as noted above, our results support a clade (Fig. 4) that includes the orders (Lampridiformes, (Polynemiformes, (Percopsiformes, (Zeiiformes, (Stylephoriformes, Gadiformes)))) and corroborate the placement of *Stylephorus chordatus* as the closest extant relative of Gadiformes (Miya et al. 2007; Chen et al. 2014; Alfar et al. 2018).

The ML phylogeny calibrated with 15 fossils suggests that Gadiformes likely originated during the Late Cretaceous, at ∼79.5 Ma (95% CI 98.0–61.0 Ma, Fig. 5; S2c), more recent than predicted by Malstrom et al. (2016). Initial divergence of crown-group Gadiformes appears to have occurred rapidly near the Cretaceous–Paleogene (K–Pg) boundary, generating three major lineages in a span of 15 million years. The earliest diverging lineage is Bregmacerotoidei (Bregmacerotidae) at ∼78.5 Ma (CI 98.0–61.0 Ma) followed by the divergence of two large lineages, which diverged at ∼71.7 Ma (CI 90.7–61.0). We tested if the recovered relationships within Gadiformes were influenced by the long branches leading to Bregmacerotidae. However, Bregmacerotidae was always recovered as the earliest branch and the ingroup relationships did not change even when the outgroups were removed, and the tree was rooted with the family. This suggests that the placement of Bregmacerotidae as sister to all remaining gadiforms is not sensitive to the systematic errors explored.

Our results consistently support four lineages following the divergence of Bregmacerotidae from other gadiforms. Gadoidei comprises four monophyletic families: Phycidae; Gadrosaridae; Lopidae; and
Figure 4. Relationships of Gadiformes based on a concatenated alignment including 58 taxa (58T-SOS-1) using a partitioned data set by codon position and best evolutionary models (P1: GTR+F+R4; P2: GTR+F+R5; P3: GTR+F+R7) with their own branch lengths to account for heterotachy. All nodes had 100% bootstrap values except when noted.
FIGURE 5. Timescale for the evolution of gadiform fishes. Phylogeny inferred for 58 taxa based on Maximum likelihood analyses of 8,244 loci (1,784,738 bp) and using a partitioned data set by codon position and best evolutionary models (P1: GTR+I+F4, P2: GTR+I+F5, P3: GTR+I+F7). Paleobiogeographic distribution of gadiform occurrences in the Paleocene and Eocene with potential migration routes of gadoids (yellow), macrouroids (white), and unknown (red) gadiform fishes according to Schwarzhans (1985, 2012), Nolf and Dockery (1993), Kriwet and Hecht (2008), Schwarzhans and Bratishko (2011), Marrama et al. 2019, and G. Carnevale (personal communication).
Gadidae. Extant members of Gadoidae, began to diverge during the Late Cretaceous to mid Eocene (69.7 Ma; CI 88.6–50.7 Ma). This clade includes one of the most intensively studied and exploited group of fishes—the commercially important Codfishes. The relationships of these fishes have not been resolved in previous morphological (e.g., Cohen 1984; Howes 1989; Endo 2002; Telethea et al. 2006; Gaemers 2017; Gaemers and Poulsen 2017) and molecular studies (e.g., von der Heyden and Matthee 2008; Roa-Varón et al. 2009; Malstrøm et al. 2016). The divergence of Ranicipitoidae, including only the monotypic Ranicipitidae (Raniceps raninus), also occurred in the Late Cretaceous to mid Paleocene (~63.2 Ma; 74.5–61 Ma). Its placement within Gadoidae based on the amino acid data set and the edge-unlinked and GHOST models should be explored by further targeted analyses. The composition of Merluccioidei is restricted to Merluccidae (including only the species of Merluccius) corroborating the original morphological concept of the family (Howes 1991, Endo 2002).

The final gadiform lineage, Macrouridei, is composed of two clades. The first clade in our preferred hypothesis (SOS-1; Macrouridea - Subclade I) is ((Euclichthyidae, (Muraenolepididae, Melanodontidae)), Trachyricinae) and has strong support (100% BS). Bootstrap values drop in SOS-3 and SOS-5 (52% and 21% BS, respectively), and the gCF and sCF percentages of decisive gene trees and sites supporting the clade were lower and the relationships changed (S2d, S17; Supplementary Material available on Dryad). Differences between the SOS-1 tree and the amino acid and summary coalescent (ASTRAL) topologies are observed in the placement of Euclichthyidae and Muraenolepididae, but relationships among these families were weakly supported in both cases (S13, S14, Supplementary Material available on Dryad). Within the second clade, Lyconus and Macronurus are the sole members of the Lyconidae and Macrouronidae, respectively, and those families form the sister clade to Bathygadidae. Steinacheria argentea, as the only member of the Steinachneridae, is the sister group of Macrouridae. The explosive diversification within Macrouridae, which gave rise to more than half of the extant gadiform species, occurred in the Late Cretaceous to mid Paleocene (~61.6 Ma; 72.5–60.0 Ma). The presence of the oldest fossil gadiforms from the Danian and Selandian of Europe and South Australia suggest a bipolar distribution by the early Paleogene, with gadoids reported in the Paleocene of North Atlantic area and North Sea Basin (Schwarzrans 2003, 2004; Kriwet and Hecht 2008; Schwarzrans and Bratiskho 2011; Schwarzrans 2012; Marramá et al. 2019) and macrourids reported in both the North Sea Basin and the Southern Ocean (Schwarzrans 1985; Kriwet and Hecht 2008; G. Carnevale, personal communication). Macrouridae traditionally have been recognized as comprising four subfamilies: Macrourinae, Bathygadinae, Trachyricinae, and Macrouroidinae (Marshall 1965; Cohen 1984; Iwamoto 1989; Nolf and Steurbaut 1989; Cohen et al. 1990; Endo 2002). Hypotheses about the interrelationships among these families often have been contentious, and the family has been regarded as likely a paraphyletic assemblage within gadiforms (Howes 1989; Roa-Varón et al. 2009). In contrast, Okamura (1989) suggested a close relationship among macrourines, trachyrincids, bathygadines, and euclichthyids. Our study supports the recognition of three macrourid subfamilies as families: Macrourinae, Bathygadinae, and Trachyrincinae (taxa included historically in both Trachyrincinae and Macrouroidinae).

This is the first study including complete taxonomic sampling at the family level and has resolved the relationships among families of Gadiformes. This phylogeny will serve as the basis for mapping molecular and morphological adaptations that might have facilitated the transition(s) between freshwater and marine environments and between shallow and deep-sea habitats. This will result in a better understanding of geographic and ecological patterns of diversification. Furthermore, this phylogenetic framework will allow the assessment of modes of colonization of deep-water habitats by fishes, which has long been debated (e.g., Andriyashev 1953, White 1988).

**Conclusions**

Our phylogenomic analyses of the teleostean order Gadiformes extensively explored potential biological and methodological sources of systematic error. We demonstrated that discarding too many loci to reduce the relative amount of missing data without considering the actual signal in the data can be detrimental. Consequently, effective gene filtering that generates optimal data matrices in terms of potential phylogenetic signal requires finding a balance between data quantity and quality. Strategies used to improve the signal-to-noise ratio (e.g., partitioning the data, estimating the most realistic model of sequence evolution, and using models addressing the rate of variation across sites and lineages) had little impact on the resulting tree topology or were insufficient to deal with all the phenomena present in our genome-scale matrices.

Changes in the tree topology for three families were observed after removing the most divergent outgroups (54T-SOS-1) and modeling rate of variation across sites and lineages (Edge-proportional partition model, Edge-unlinked partition model, and GHOST), suggesting that theoretical work is needed to explore the relationship between outgroup rooting and LBA, especially how different outgroup taxon sampling strategies affect the probability of LBA.

Using multiple approaches for assessing phylogenetic confidence, such as gCF and sCF analyses, serves to complement measures such as bootstrap values by capturing variance in phylogenetic signal and identifying hidden conflict at different evolutionary scales. Bootstrap contraction thresholds (e.g., 20%)
recovered the monophyly of Lotidae and lead to greater accuracy in species tree inference using ASTRAL-III. Most of the gene-tree and species-tree discordance is a result of short times between divergences and consequent lack of informative characters at deep levels rather than a result of ILS.

This study provides a higher-level classification that is of operative value for a wide range of comparative studies of multiple traits in an intensively harvested group of fishes. The highly resolved backbone of our phylogeny alters historical perceptions of evolution within the group. Changes in classification are necessary and a revised classification comprising five suborders and 17 families is proposed. While not the final effort on inferring higher-level gadiform relationships, our analyses have full representation at the family level, including a number of difficult-to-obtain specimens (e.g., Euclichthyidae, Lyconidae, Ranicipitidae) that have not previously been included in a molecular study. In addition, this is the first explicitly time-calibrated tree using 15 fossil calibration points for Gadiformes, and the dating analyses indicate that the order probably originated in the North Atlantic and diversified rapidly in the late Cretaceous, a result confirmed by recent paleontological discoveries (G. Carnevale, personal communication).

The large amount of congruence across analyses and sequence-bias testing increases confidence in the results across different evolutionary depths. Future efforts will focus on refining the bait set to mask over-sequenced regions (evening read coverage across loci) and increasing taxonomic sampling within each family. Finally, our results reiterate the importance of examining phylogenomic data sets for evidence of systematic error that can emerge as a result of unsuitable modeling of biological and/or methodological issues, even when data sets are large and yield high support for phylogenetic relationships.

SUPPLEMENTARY MATERIAL

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.5qfttdz2w.

ACKNOWLEDGMENTS

We thank the following colleagues who kindly provided tissue samples for our study: Tomio Iwamoto, Peter McMillan, Sophie von der Heyden, Hiroimitsu Endo, Rob Leslie, Steve W. Ross, Rafael Bañon, Luis M. Adasme, and Juan M. Diaz de Astráloa. The computations performed for this article were conducted on the Smithsonian High-Performance Cluster (SI/HPC), Smithsonian Institution. https://doi.org/10.25572/SIHPC. A.R-V. thanks W. Wang, Q. Wang, S. Song, X. Zheng, L. Peng, and J. Jiang for all their support in her time at Shanghai Ocean University. This research benefited from data and technical support provided by the Smithsonian bioinformatics working group, including Matt Kweskin, Vanessa Gonzalez, Mirian Tsuchiya, and Mike Trizna. We thank Tomio Iwamoto, Amy Driskell, Noor White, Lynne R. Parenti, and Bruce Collette for constructive comments on this manuscript. A.R-V. further thanks Allen G. Collins and Naikoa Aguilar-Amuchastegui for their continuous support and encouragement. This is contribution number 3953 of the Virginia Institute of Marine Science, William and Mary.

FUNDING

Funding included National Science Foundation (NSF) East Asia and Pacific Summer Internship (Award number 1514994 to A.R-V.) to conduct laboratory work in Shanghai Ocean University, NSF DEB-1601433 (DDIG to E.J.H. for A.R-V.), and a Smithsonian Institution Predoctoral Fellowship to A.R-V. Additional funding was provided by the Virginia Institute of Marine Science.

COMPETING INTERESTS

The authors have declared that no competing interests exist.

AUTHOR’S CONTRIBUTIONS

A.R-V. designed the experiments, carried out the lab work, designed and performed the data analyses, made all figures, and wrote the manuscript. R.B.D. participated in the data analyses and reviewed the final manuscript. G.C. selected appropriate calibration points and reviewed the final manuscript. L.T. provided support at an early stage of the Trinity assembly and reviewed the final manuscript. C.C.B. reviewed and edited multiple versions of the manuscript and supported A.R-V.’s pre-doctoral fellowship at the Smithsonian Institution, NMNH. C.L. provided training in the molecular approach to A.R-V. in his lab in Shanghai, pipeline and scripts associated with it, and technical support provided by the Virginia Institute of Marine Science.

REFERENCES


Schwarzshans W., Stringer G. L. 2020. Fish otoliths from the late Maastrichtian Kemp Clay (Texas, USA) and the early Danian Clayton Formation (Arkansas, USA) and an assessment of extinction and survival of teleost lineages across the K-Pg boundary based on otoliths. Riv. It. Paleont. Strat. 126:395–446.


