Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data

Lily C. Hughes[a,b,1,2], Guillermo Ortí[a,b,1,2], Yu Huang[c,d,1], Ying Sun[c,e,1], Carole C. Baldwin[b], Andrew W. Thompson[a,b], Dahiana Arcila[a,b], Ricardo Betancur-R.[b,f], Chenhong Li[g], Leandro Becker[h], Nicolás Bellora[h], Xiaomeng Zhao[c,d], Xiaofeng Li[c,d], Min Wang[c], Chao Fang[d], Bing Xie[c], Zhuocheng Zhou[i], Hai Huang[j], Songlin Chen[k], Byrappa Venkatesh[l,2], and Qiong Shi[c,d,2]

[a]Department of Biological Sciences, The George Washington University, Washington, DC 20052; [b]National Museum of Natural History, Smithsonian Institution, Washington, DC 20560; [c]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, Beijing Genomics Institute Academy of Marine Sciences, Beijing Genomics Institute Marine, Beijing Genomics Institute, 518083 Shenzhen, China; [d]Beijing Genomics Institute Education Center, University of Chinese Academy of Sciences, 518083 Shenzhen, China; [e]China National GeneBank, Beijing Genomics Institute-Shenzhen, 518120 Shenzhen, China; [f]Department of Biology, University of Puerto Rico–Rio Piedras, San Juan 00931, Puerto Rico; [g]Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Shanghai Ocean University, Ministry of Education, 201306 Shanghai, China; [h]Laboratorio de Ictiología y Acuicultura Experimental, Universidad Nacional del Comahue–CONICET, 8400 Bariloche, Argentina; [i]Professional Committee of Native Aquatic Organisms and Water Ecosystem, China Fisheries Association, 100125 Beijing, China; [j]College of Life Science and Ecology, Hainan Tropical Ocean University, 572022 Sanya, China; [k]Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, 266071 Qingdao, China; and [l]Comparative Genomics Laboratory, Institute of Molecular and Cell Biology, A*STAR, Biopolis, 138673 Singapore

Edited by Scott V. Edwards, Harvard University, Cambridge, MA, and approved April 13, 2018 (received for review November 7, 2017)

Our understanding of phylogenetic relationships among bony fishes has been transformed by analysis of a small number of genes, but uncertainty remains around critical nodes. Genome-scale inferences so far have sampled a limited number of taxa and genes. Here we leveraged 144 genomes and 159 transcriptomes to investigate fish evolution with an unparalleled scale of data: >0.5 Mb from 1,105 orthologous exon sequences from 303 species, representing 66 out of 72 ray-finned fish orders. We apply phylogenetic tests designed to trace the effect of whole-genome duplication events on gene trees and find paralogy-free loci using a bioinformatics approach. Genome-wide data support the structure of the fish phylogeny, and hypothesis-testing procedures appropriate for phylogenomic datasets using explicit gene genealogy interrogation settle some long-standing uncertainties, such as the branching order at the base of the teleosts and among early euteleosts, and the sister lineage to the acanthomorph and percomorph radiations. Comprehensive fossil calibrations date the origin of all major fish lineages before the end of the Cretaceous.

phylogenomics | divergence times | bony fish | paralogy | tree of life

Ray-finned fishes have evolved over more than 400 million years to occupy aquatic environments worldwide, from deep ocean trenches to high mountain streams, and thrive in extreme habitats with acidic, subzero, hypersaline, hypoxic, temporary, and fast-flowing water conditions (1). Establishing their phylogenetic relationships is a fundamental step toward unraveling the evolutionary processes responsible for this diversity. Knowledge of the phylogeny of fishes is far from complete but has significantly advanced in recent years, especially by identifying major lineages within the vast diversity of percomorphs (>17,000 species) through analysis of ∼20 gene fragments (2, 3). These comprehensive molecular phylogenies upended classical morphological hypotheses and inspired the synthesis of a new phylogenetic classification, with consensus gradually developing among ichthyologists (4). However, these major, relatively recent discoveries for the backbone of the fish tree of life have not been tackled by comprehensive, genome-scale datasets (but see refs. 5–9 for taxonomically restricted studies). Particular areas of contention remain, involving the branching order at the base of the teleosts, the relationships among otophysan orders, the sister group to the neoteleosts, the sister group to the acanthopterygians, and the relationship among atherinomorph orders.

Currently established high-throughput sequencing technologies enable systematists to analyze hundreds to thousands of loci for phylogenetic analysis (10–12). However, big datasets present new methodological challenges to defining the most informative orthologous loci and appropriate tree-inference methods (13). In fishes, these challenges are further compounded by the incidence of duplicate genes resulting from two whole-genome duplications (WGDs) among early vertebrates (VGD1 and VGD2) and a WGD event in the common ancestor of teleosts (TGD) (Fig. 1), in addition to more recent lineage-specific WGDs in some fishes (14–16). Distinguishing

Significance

Ray-finned fishes form the largest and most diverse group of vertebrates. Establishing their phylogenetic relationships is a critical step to explaining their diversity. We compiled the largest comparative genomic database of fishes that provides genome-scale support for previous phylogenetic results and used it to resolve further some contentious relationships in fish phylogeny. A vetted set of exon markers identified in this study is a promising resource for current sequencing approaches to significantly increase genetic and taxonomic coverage to resolve the tree of life for all fishes. Our time-calibrated analysis suggests that most lineages of living fishes were already established in the Mesozoic Period, more than 65 million years ago.

Author contributions: L.C.H., G.O., Y.H., Y.S., C.C.B., R.B.-R., B.V., and Q.S. designed research; L.C.H., G.O., Y.H., A.W.T., D.A., R.B.-R., L.B., N.B., X.Z., X.L., M.W., C.F., B.X., and Q.S. performed research; L.C.H., G.O., Y.H., C.C.B., A.W.T., D.A., R.B.-R., Z.Z., H.H., S.C., B.V., and Q.S. contributed new reagents/analytic tools; L.C.H., G.O., Y.H., A.W.T., D.A., R.B.-R., C.L., L.B., N.B., X.Z., X.L., M.W., C.F., and B.X. analyzed data; and L.C.H., G.O., Y.H., D.A., R.B.-R., B.V., and Q.S. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The accession numbers reported in this paper are listed in *SI Appendix*, Table S1, and have been deposited to the National Center for Biotechnology Information (NCBI) BioProject (https://www.ncbi.nlm.nih.gov/bioproject/398732). Additional analysis files have been deposited to the Dryad Digital Repository (https://doi.org/10.5061/dryad.5b85783).

See Commentary on page 6107.

[1]L.C.H., G.O., Y.H., and Y.S. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: lilychughes@gwu.edu, gorti@gwu.edu, mcbbv@imcb.a-star.edu.sg, or shiqiong@genomics.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1719358115/-/DCSupplemental.

Published online May 14, 2018.

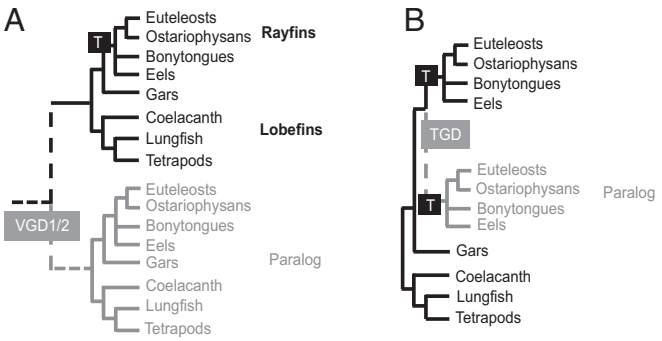

Fig. 1. Effects of WGDs on individual gene trees and the inference of organismal phylogeny. Gene genealogies resulting from (*A*) the VGD1/2 event showing duplicate (paralogous) groups of rayfins and lobefins, and (*B*) the TGD event showing duplicated clades of euteleosts and ostariophysans. These known WGDs are misleading for organismal phylogeny, but provide testable hypotheses to identify paralogs.

orthologous genes, whose sequence divergence follows speciation, from duplicated paralogs is a crucial task for resolving the tree of life (17). However, the popular set of exons successfully used for fish phylogenetics under the PCR-Sanger sequencing paradigm (18) and newer genomic marker sets obtained by target capture (7) have not been explicitly tested for orthology. Genetic markers in use are normally selected by screening one or a few genomes available for model organisms, without applying stringent analyses to determine orthology beyond similarity-based criteria (7, 18, 19).

To expand currently available genomic resources for fishes, we established an international consortium (20) to sequence transcriptomes sampled from a wide taxonomic diversity (https://db.cngb.org/fisht1k/). Transcriptomic data for 131 fish species combined with whole-genome sequences of fishes (21, 22) offered an unprecedented opportunity for establishing a genome-wide set of loci for phylogenomic analysis. While many types of coding and noncoding genomic loci such as ultraconserved elements (UCEs), introns, or conserved nonexonic elements (CNEEs) are available for reconstructing phylogeny (19), we focused on exons to significantly expand the existing databases that have been used successfully to reconstruct fish phylogeny at large scales (2, 3). Exons are easy to align, may be partitioned by codon position, and can be translated to amino acids to minimize artifacts from base compositional biases (23). After identifying a set of verifiably orthologous exon markers, we compiled the largest phylogenomic matrix for fishes to date (*SI Appendix*, Fig. S1) and inferred a phylogeny for 300 actinopterygian taxa and three outgroups, representing all major ray-finned fish lineages. Additionally, we used an explicit hypothesis-testing approach to interrogate six areas along the backbone phylogeny with controversial relationships (Fig. 2) and time calibrated the tree with comprehensive fossil data to test current hypotheses of teleost diversification (5, 24).

Results

Orthology Assessment. We explicitly addressed the issue of orthology in exon markers in the context of multiple WGDs based on a comprehensive dataset of transcriptomes and recently released whole-genome sequences (*SI Appendix*, Table S1). Using the EvolMarkers pipeline (25) applied to eight well-annotated genomes distributed across the fish tree of life (Fig. 2, asterisks and *Materials and Methods*), we identified a set of 1,721 single-copy, conserved exons (>60% identity among taxa) longer than 200 bp. Exon alignments for these eight model species were used to parameterize hidden Markov models (HMMs) to search for each locus in 144 genomes and 159 transcriptomes (*SI Appendix*, Table S1). Significant hits were added to each original alignment to create an exon sequence database for 303 species to subsequently test for orthology. Individual gene trees were inferred for all loci and analyzed to detect gene duplications at the base of the vertebrate tree (VGD1/2) and at the base of the teleosts (TGD). A subset of 469 paralogous genes explained by the VGD1/2 events was identified via topology tests (26) on the monophyly of teleosts, and excluded from further analysis (Fig. 1*A*). We also excluded 111 paralogous loci originating from the TGD event after testing for euteleost monophyly, and 36 additional loci after testing for ostariophysan monophyly (Fig. 1*B*). After explicitly accounting for WGD events in our thorough assessment of orthology with comprehensive taxonomic sampling, we excluded 616 loci to define a final set of 1,105 paralogy-free exons.

Structure of the Ray-Finned Fish Tree of Life. The concatenated alignment of 1,105 exons produced a data matrix with 555,288 bp (185,096 amino acids) for 303 species. We used protein translations for phylogenetic analysis to reduce the confounding effect of base-composition heterogeneity among taxa, a biasing factor shown to be extensive among fishes (23), and resolved the backbone phylogeny with confidence (Fig. 2). Maximum-likelihood (ML) analyses of concatenated nucleotide and protein sequences, Bayesian analysis of proteins, and a summary multispecies coalescent approach converged on virtually the same topology (*SI Appendix*, Figs. S2–S5), with some conflicting nodes discussed below. The average bootstrap support values for the ML trees is 94%. Our phylogeny based on significantly more genetic loci relative to previous studies is remarkably congruent with other analyses of actinopterygians before the advent of genomic datasets (2, 3). By contrast, phylogenetic results obtained with the set of 616 putatively paralogous loci detected by our testing procedure contained several unconventional groupings and resulted in the nonmonophyly of well-established taxa, especially among percomorphs (*SI Appendix*, Fig. S6).

Tests of Alternative Hypotheses. Confidence in phylogenetic results requires not only large amounts of data and high bootstrap values but also assessment of conflicting phylogenetic signal among gene trees. To gauge the extent of incongruence present in our phylogenomic data matrix, we calculated tree certainty (TC) for ML trees based on nucleotide and protein sequences (27, 28). Relative TC values ranged from 0.35 to 0.43, suggesting a low level of incongruence among gene trees overall, in agreement with high bootstrap support, and congruent topologies were obtained for the backbone of the trees for all analyses. However, some critical nodes resolved by our analyses mask highly conflicting gene-tree distributions (Fig. 2). Internode certainty (ICA) values for edges subtending these nodes are close to zero or negative (*SI Appendix*, Table S2). For these cases, we applied explicit gene genealogy interrogation (GGI) to test alternative hypotheses (9, 29) (Fig. 3 and *SI Appendix*, Fig. S7). Individual gene trees are extremely useful for reconstructing the species tree while taking into account coalescent variation that might mislead a concatenated analysis. However, since individual gene alignments can be short and hold insufficient information, trees inferred from these genes are prone to error. GGI is based on topology tests to identify the genealogical history, among a set of predefined alternatives, that each gene supports with highest probability. Selection of a preferred hypothesis by GGI discerns between actual genealogical incongruence and estimation error arising from limited signal in short sequence alignments. We also calculated the recently proposed ΔGLS metric (30) between the top two competing hypotheses. This comparison shows that the hypotheses preferred by GGI also have better likelihood scores on a gene-by-gene basis than the alternative topologies for a majority of genes (*SI Appendix*, Fig. S8).

Contrary to the prevailing morphological view of basal teleostean relationships (31), our concatenated and multispecies coalescent analyses support Elopomorpha (eels, tarpons) and Osteoglossomorpha (bony tongues, mooneyes) as sister taxa (Fig. 2*A*). This hypothesis has been supported before by molecular studies (32),

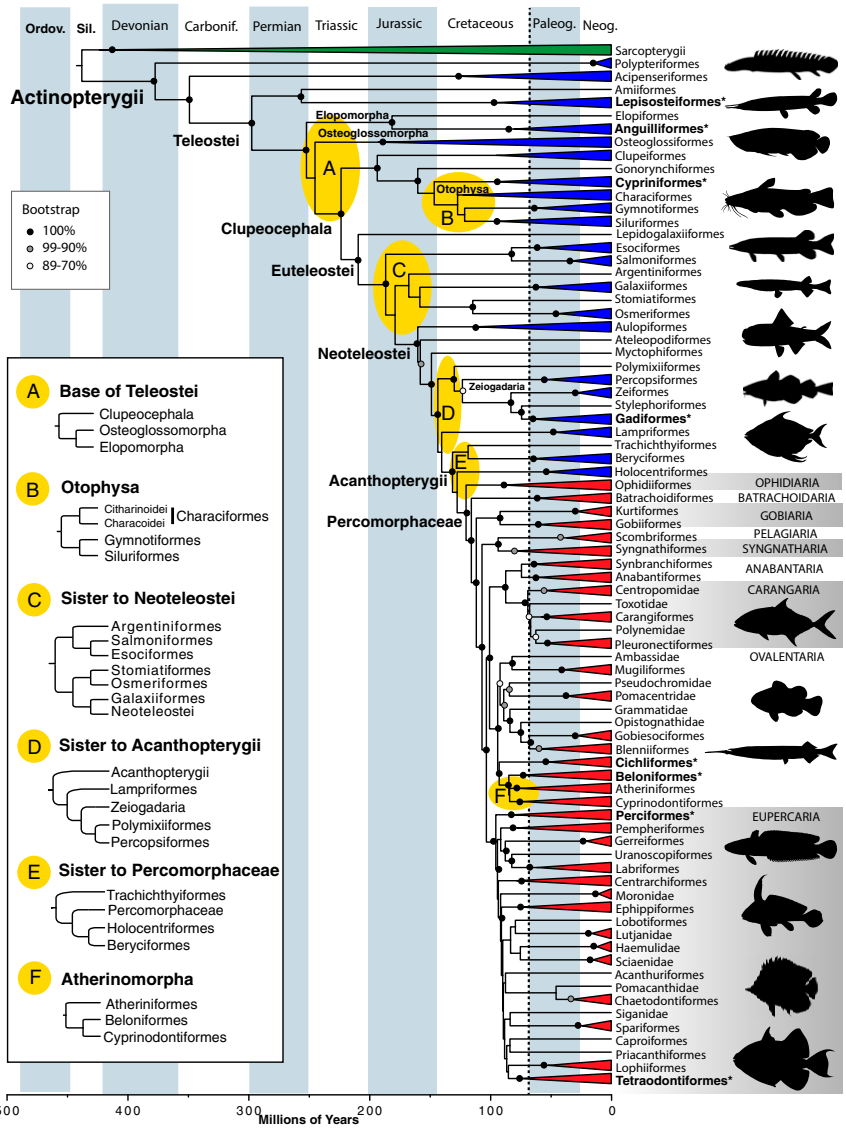

Fig. 2. Fossil-calibrated phylogeny of 300 actinopterygians (blue and red) and three sarcopterygian outgroups (green) with divergence dates estimated in MCMCTree (see *SI Appendix*, Fig. S9 for the uncollapsed tree). The percomorph radiation is highlighted in red, with the names of the nine percomorph series (*Right*). The topology of the chronogram reflects the ML tree inferred from 185,096 amino acid sites, except where GGI strongly preferred an alternative topology for the areas highlighted in yellow ovals (*A*–*F*). (*Left*) ML topology for *A*, *C*, *E*, and *F*. For *B*, "Otophysa," the GGI and ML protein topologies agree, supporting a monophyletic Characiformes, although there is discordance with the nucleotide results (*SI Appendix*, Figs. S2 and S5). For *D*, the GGI nucleotide results agree with the ML concatenated trees placing Lampriformes as the sister to Acanthopterygii, but an alternative hypothesis has some support with GGI protein results and is shown (*Left*) (Fig. 3*E*). Taxa marked with asterisks contained one of the eight model genomes. The dotted vertical line denotes the Cretaceous–Paleogene boundary. Animal silhouettes courtesy of PhyloPic (http://www.phylopic.org).

but GGI tests clearly favor placing elopomorphs alone as the sister group to all other teleosts (Fig. 3*A*). Relationships among otophysan orders also are influenced by high levels of conflicting phylogenetic signal (*SI Appendix*, Table S2). GGI tests support the canonical hypothesis proposed by morphology (Figs. 2*B* and 3*B*) (33), a result that has been elusive for previous molecular studies (34) but was recently resolved using GGI applied to a different phylogenomic dataset (9). GGI tests overwhelmingly support Holocentriformes (squirrelfishes, soldierfishes) as the closest living relatives of the percomorphs (Fig. 3*E*), rejecting the topology obtained via analyses of protein and nucleotide sequence data (*SI Appendix*, Figs. S2–S5). Similarly, GGI resolves the relationship among the three orders of Atherinomorpha (Fig. 3*F*) supporting a clade composed of Cyprinodontiformes (killifishes, mollies, guppies) and Atheriniformes (silversides, rainbowfishes), to the exclusion of Beloniformes (ricefishes, flying fishes, halfbeaks).

Discussion

Phylogenetic resolution of relationships among groups of fishes has lagged behind other vertebrate groups, partly because until recently, there was a paucity of genomic resources for fishes. By generating 131 transcriptomes, we have expanded these resources significantly. Previous molecular phylogenies of fishes established many novel clades never before identified on the basis of morphological evidence (2, 3), but occasionally with weak support. Our massive dataset for all major ray-finned fish lineages corroborates the recently proposed phylogenetic classification of bony fishes with major groups such as teleosts and successively branching clupeocephalans, euteleosts, neoteleosts, and acanthomorphs, leading to the most species-rich clade of modern fishes, the percomorphs (Fig. 2) (4) (see *SI Appendix*, Table S3 for node-by-node comparisons with previous studies).

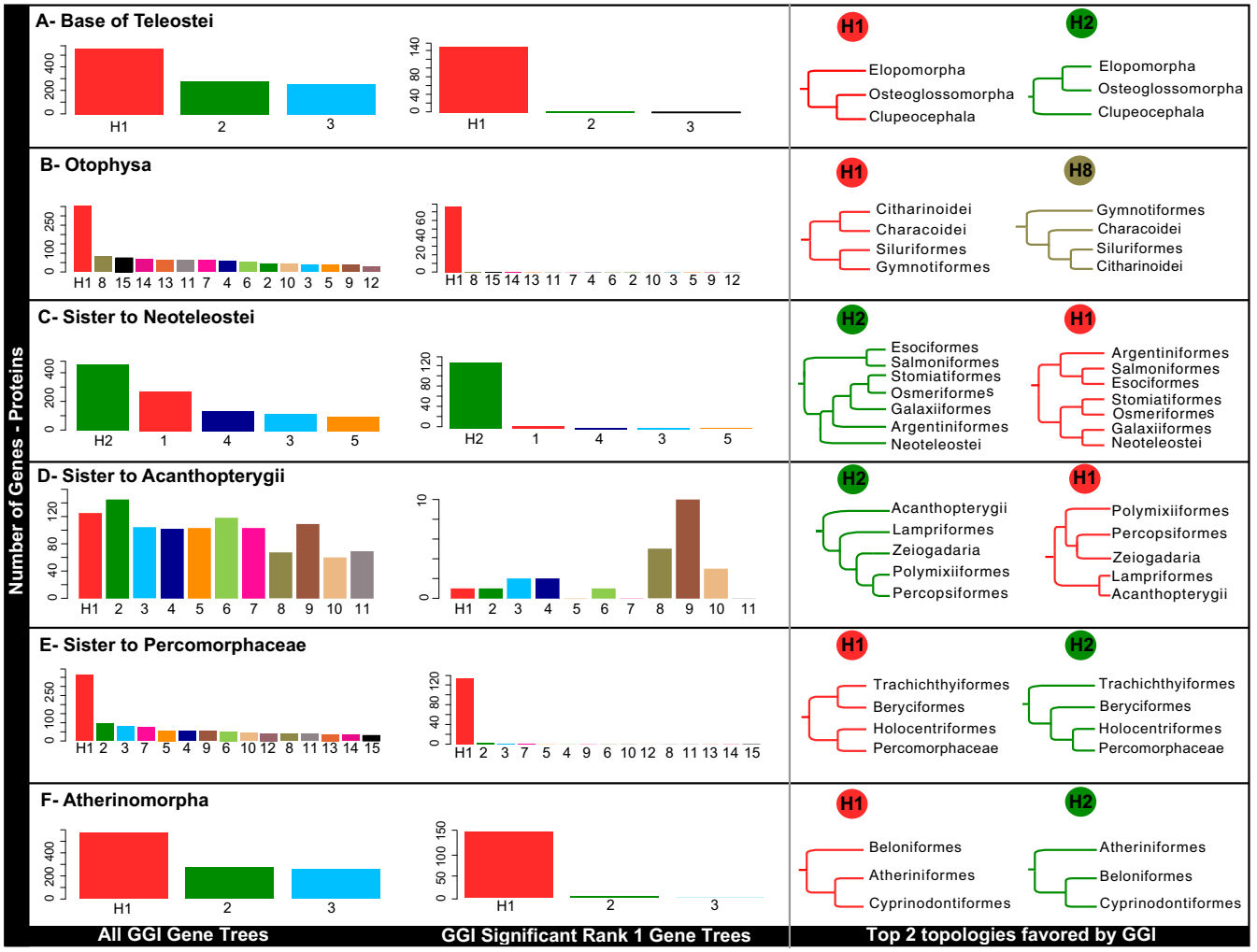

Fig. 3. GGI results based on protein alignments. For each specific hypothesis tested (*A–F*), the distribution of all gene trees supporting each alternative hypothesis (*Left*) or only the significantly supported hypotheses (*Middle*) are shown. The top two topologies favored by these tests are shown on the *Right* and in Fig. 2. GGI results based on nucleotide alignments are shown in *SI Appendix*, Fig. S7.

The mean estimate for the age of crown actinopterygians was ~379 Ma (Fig. 2 and *SI Appendix*, Fig. S9), 20 million years older than a recent estimate obtained after reassignment of older actinopterygian fossils (24).

The subdivision of percomorphs into nine series was resolved with high support: 99 to 100% bootstrap values for both protein and nucleotide sequence analyses (Fig. 2 and *SI Appendix*, Figs. S2 and S3). We obtain a sister-group relationship between the clade containing the fastest-swimming species in Pelagiaria (tunas) and the more languid species in Syngnatharia (seahorses, pipefishes). Another group receiving strong support includes a set of phenotypically disparate taxa within Carangaria, harboring extremely asymmetric benthic flatfishes (Pleuronectiformes) together with jacks (Carangiformes) and other strong pelagic swimmers like the dolphinfish, as the sister to Anabantaria, containing many air-breathing and predominantly freshwater fishes (Fig. 2). We find strong support for the series Ovalentaria (35), composed of cichlids, atherinomorphs, blennies, clingfishes, and other allies notable for their sticky egg filaments (though lost in some lineages). These results also support the circumscription of Perciformes as an order within Eupercaria that includes (among others) basses, groupers, sticklebacks, and sculpins (3, 4), though many more taxa are needed to resolve the hyperdiverse Eupercaria clade of more than 163 families. The macroevolutionary pattern revealed by our results (Fig. 2) is congruent with substantial percomorph diversification occurring before the end of the Cretaceous, with major lineages already established well before the Cretaceous–Paleogene (K–Pg) extinction event (36), remarkably surviving the mass extinction event that decimated the dinosaurs. Unlike divergence estimates obtained with UCEs (5), our results do not support explosive diversification of percomorphs (especially Eupercaria and Ovalentaria) at the K–Pg boundary. In contrast, rapid radiation of these groups is inferred to occur in the Late Cretaceous (Fig. 2 and *SI Appendix*, Fig. S9).

A few areas of the backbone phylogeny remain enigmatic and probably require additional taxonomic sampling to achieve better resolution. Furthermore, systematic error in large concatenated datasets may be exacerbated when major model assumptions (such as base compositional stationarity) are not met (23), and gene-tree estimation error is known to mislead summary coalescent approaches (37). These factors may be compounded by missing data in poorly sampled clades (38). Our analyses identified several areas (Fig. 3 *C* and *D*) that require careful consideration. First, the position of Argentiniformes and Galaxiiformes in relation to neoteleosts and other protacanthopterygian lineages (Fig. 2*C*) is undermined by high gene-tree conflict, with conflicting topologies from concatenated ML protein and nucleotide analysis (*SI Appendix*, Figs. S2 and S3 and Table S2). Sparse taxonomic sampling in this area of our phylogeny yields ambiguous results under GGI

(Fig. 3*C* and *SI Appendix*, Fig. S7*C*). Second, the placement of Lampriformes and Polymixiiformes in relation to the acanthopterygians varies in this study, since both concatenation ML inference and multispecies coalescent approaches resolve Lampriformes as the sister group to acanthopterygians but Bayesian inference supports an alternative topology (*SI Appendix*, Fig. S4). GGI tests support Lampriformes as the sister to Acanthopterygii for nucleotides (*SI Appendix*, Fig. S7*D*) but proteins fail to differentiate between hypotheses, with minor support for Lampriformes forming a clade with other paracanthopterygian lineages as was recently obtained with UCEs (Fig. 3*D*) (5).

Paralogs in Phylogenomic Datasets. The history of vertebrate WGDs generates testable hypotheses for detecting paralogy in gene trees (Fig. 1), but this is rarely taken into account during phylogenetic analysis. We find that orthology assessment based on only a few model genomes significantly underestimates the number of paralogs. The transcriptomes sequenced here, and the many draft genome sequences generated recently for fishes, provide a rich database for orthology assessment using explicit hypothesis testing based on gene trees. As additional genomes and transcriptomes become available in the future, our ability to detect paralogy will increase.

Many widely used exon markers for fish phylogenetics were originally developed by screening just two model genomes available a decade ago for PCR-based sequencing (18). However, four exons, RAG1, RAG2, GLYT, and FICD, for which many sequences are already available, were deemed paralogy-free. With more than 17,000 fish sequences for RAG1 available in GenBank (as of May 17, 2017), these public sequences can be integrated with newly generated data from current sequencing technology. Our results also show that the specificity of PCR primers used in previous studies avoided the amplification of paralogous genes (*SI Appendix*, Figs. S10–S12), and therefore previous results using TBR1, MYH6, and KIAA1239 markers were not compromised by hidden paralogy. However, current target-capture technology, now in wide use in phylogenomics, does not rely on PCR primers but rather on single-stranded RNA probes that hybridize with genomic DNA, producing many short reads from targeted areas of the genome (39). Assembling these short reads under the assumption that they come from a single ortholog is potentially problematic, as even a few stray reads from a paralogous locus could create a chimeric assembly, violating phylogenetic assumptions and introducing error into the analysis. By avoiding loci with known paralogs, assembly and analysis of target-capture data are greatly simplified.

Exon Markers for Future Phylogenomics. Phylogenomic analyses do not necessarily require complete genomes or transcriptomic datasets. Cost-effective, reduced-representation approaches that target a sufficient number of carefully selected genetic loci hold promise to accurately resolve phylogeny by minimizing estimation error and maximizing taxonomic sampling (39). We provide a set of markers that can easily be sequenced through massive sequence-capture experiments that also connect with older PCR-based datasets and new genomic resources, unlike recently developed markers based on UCEs that demand de novo sequencing efforts and a posteriori assessment of paralogy (10). Other promising noncoding markers such as CNEEs (19) or introns may provide independent evidence to resolve recalcitrant clades, but introns are too variable to align for deeply diverging lineages and the utility of CNEEs in fishes may be diminished due to the extraordinary high rate of molecular evolution and loss of noncoding elements in teleost genomes relative to tetrapods (40). The set of orthologous exons identified here will enable large-scale phylogenomic studies with dense sampling of fish taxa, while avoiding potential problems with paralogy that confound the assembly of captured loci. Thus, resolution of the tree of life for all fishes is well within our reach.

Materials and Methods

Taxonomic Sampling and RNA Sequencing. A total of 131 transcriptomes included in this study were newly sequenced as a part of the Transcriptomes of 1,000 Fishes Project (Fish-T1K) (20). Specimens representing all major lineages of fishes were collected in the field by numerous colleagues around the world (*SI Appendix*, Table S1). Protocols were reviewed and approved by the Institutional Review Board on Bioethics and Biosafety of Beijing Genomics Institute (BGI) (BGI-IRB 15139). Transcriptomes were sequenced at BGI-Tech. Sequencing details can be found in *SI Appendix*, *Materials and Methods*.

Marker Selection. An initial set of 17,817 single-copy conserved nuclear coding markers was identified by Song et al. (12) by comparing eight well-annotated fish genomes: *Lepisosteus oculatus*, *Anguilla anguilla*, *Danio rerio*, *Gadus morhua*, *Oryzias latipes*, *Oreochromis niloticus*, *Gasterosteus aculeatus,* and *Tetraodon nigroviridis* (hereafter referred to as the "model species"). We reduced the original set down to 1,721 exon markers >200 bp. To search for these loci in the transcriptomic and genomic datasets ("nonmodel species"), we used the HMM approach available in HMMER 3.1 (41). For each of the 1,721 eight–model-species alignments, we parameterized an HMM profile and executed an nHMMER search on each of our nonmodel genomes and transcriptomes (*SI Appendix*, Table S1), using default settings. Zero, one, or more hits were obtained for each marker for each of the nonmodel species for subsequent analyses and extracted from genomes and transcriptomes with custom Python scripts (available on Dryad; *SI Appendix*, *Appendix 1*). Maximum-likelihood analyses were conducted in RAxML v8.2.9 (42) for each DNA alignment, yielding 1,721 unconstrained ML gene trees, one for each exon marker.

Paralogy Filtering. Two sets of topological constraints were generated and analyzed to test hypotheses of paralogy originating from inferred WGDs in ancestral vertebrates or teleosts (Fig. 1*A*). The first constrained tree enforced the monophyly of all teleost sequences in a given gene alignment. Topology tests were subsequently performed comparing the constrained tree with the unconstrained ML topology with the expectation that teleost monophyly would be rejected due to the presence of duplicated loci originating from the ancestral vertebrate WGD events (VGD1/2; Fig. 1). Ten separate unconstrained ML gene-tree searches and tree searches constraining the monophyly of teleosts for each locus were conducted in RAxML under the GTRGAMMA model. A second set of ML gene trees was inferred by enforcing monophyly of Ostariophysi (Fig. 1), a large and well-supported clade within teleosts (3, 9), to identify paralogs originating from the TGD. For the topology tests, we obtained site likelihoods using RAxML, and then applied the approximately unbiased (AU) test in Consel (26, 43) to evaluate whether constrained topologies could be rejected for each gene tree compared with the unconstrained topology for the same locus. Rejection of teleost monophyly (AU test, $P < 0.05$) was used to flag the marker as a potential paralog originating from the VGD1/2 events. Loci that passed this step were tested for the potential effects of the TGD, this time comparing ostariophysan monophyly-constrained trees against the unconstrained tree for each locus. We also tested the monophyly of a third group, the euteleosts (Fig. 1), for which taxon sampling was more complete than for ostariophysans, as an independent test of the TGD. When monophyly of any of these groups was significantly rejected by the AU test, the locus alignment was considered to contain paralogs and was removed from further phylogenetic analysis. Seven exon markers commonly used for PCR-based studies of fish phylogenetics (18), TBR1, FICD, RAG1, RAG2, GLYT, KIAA1239, and MYH6, were also tested for paralogy as described above, and flagged loci were analyzed further to test whether the published data (sequences obtained by PCR-Sanger sequencing methods) may contain paralogous sequences, potentially confounding previous studies (*SI Appendix*, *Materials and Methods*).

Phylogenomic Analyses and Gene Genealogy Interrogation. A species tree was inferred using ASTRAL-II (44) with individual RAxML-estimated gene nucleotide trees, with each locus alignment partitioned by codon and optimized under the GTRGAMMA model. Concatenated protein and nucleotide analyses were conducted in ExaML (45), and a Bayesian search was conducted on proteins in ExaBayes (46). We also calculated internode certainty (IC and ICA or IC All) and tree certainty (TC and TC All or TCA) values in RAxML for the ML concatenated protein tree, using both the protein and nucleotide gene trees (28, 47). Additionally, to address the effect of our paralogy filtering,

we analyzed a DNA matrix consisting of the 616 loci discarded by our filter, using the best hit from the nHMMER search for a particular locus. We conducted relaxed-clock divergence time estimates in MCMCTree (48), with more than 30 fossil calibrations with a subset of our data of the 21 most complete loci (10,203 bp), for which both runs converged with effective sample size values >200 (*SI Appendix, Materials and Methods*).

We used GGI to test long-standing areas of conflict in the ray-finned fish tree of life, shown in yellow circles in Fig. 2. The details of this method are described in full by Arcila et al. (9), but briefly, we generated topological constraints for each hypothesis we tested (available on Dryad), and then conducted 10 ML gene-tree searches for each of those constrained topologies in RAxML using both the nucleotide and protein alignments (Fig. 3 and *SI Appendix*, Fig. S10). We calculated site likelihoods for all ML constrained tree topologies for each locus in RAxML and used the AU test in Consel to rank the best-supported topology. Details on the specific hypotheses tested and relevant references can be found in *SI Appendix, Materials and Methods* and Fig. S13. Additional files are available at Dryad Digital Repository, https://doi.org/10.5061/dryad.5b85783.

ACKNOWLEDGMENTS. We thank all Fish-T1K partners. Special thanks go to our colleagues who provided samples: Yong Zhang (Sun Yat-Sen University), Junxing Yang (Kunming Institute of Zoology), Jun Zhao (South China Normal University), Jose V. Lopez and Kirk Kilfoyle (Nova Southeastern University), Brad Pusey and Stephen Beatty (James Cook University), Peter Rask Møller and Mads Reinholdt (Natural History Museum of Denmark), Luiz Rocha (California Academy of Sciences), Donald J. Stewart (SUNY College of Environmental Science and Forestry), Carol A. Stepien (NOAA-PMEL), and Andrew I. Furness (University of California, Irvine). Thanks to Yong Zhang, Zhixiang Yan, Shanshan Liu, Yamin Huo, Jing Zhang, and Nanxi Liu (China National GeneBank, BGI-Shenzhen) for their efforts in sample processing and data management. Yadira Ortiz-Ruiz (University of Puerto Rico) and Robert Kallal and Jesus Ballesteros (The George Washington University) helped with data analysis. Thanks to Colonial One high-performance computing cluster personnel (The George Washington University). We thank W. Leo Smith, Keith Crandall, Vanessa Gonzalez, Elizabeth Alter, and R. Alexander Pyron, and two anonymous reviewers for their comments, which greatly improved the quality of this manuscript. This work was supported by funding from the National Natural Science Foundation of China (31370047) and Shenzen Special Program for Strategic Emerging Industries (JSGG20170412153411369) (to Q.S.), Biomedical Research Council of Singapore and A*STAR (B.V.), grants from the US National Science Foundation (DEB-147184 and DEB-1541491, to R.B.-R.; DEB-1457426 and DEB-1541554, to G.O.), a grant from The Smithsonian Institution (GGI-SIBG Awards Program) (to C.C.B.), and additional support from the Harlan Endowment to The George Washington University.

1. Helfman GS, Collette BB, Facey DE, Bowen BW (2009) *The Diversity of Fishes* (John Wiley & Sons, Oxford), 2nd Ed.
2. Near TJ, et al. (2012) Resolution of ray-finned fish phylogeny and timing of diversification. *Proc Natl Acad Sci USA* 109:13698–13703.
3. Betancur-R R, et al. (2013) The tree of life and a new classification of bony fishes. *PLoS Curr* 5:ecurrents.tol.53ba26640df0ccaee75bb165c8c26288.
4. Betancur-R R, et al. (2017) Phylogenetic classification of bony fishes. *BMC Evol Biol* 17: 162.
5. Alfaro ME, et al. (2018) Explosive diversification of marine fishes at the Cretaceous-Palaeogene boundary. *Nat Ecol Evol* 2:688–696.
6. Longo SJ, et al. (2017) Phylogenomic analysis of a rapid radiation of misfit fishes (Syngnathiformes) using ultraconserved elements. *Mol Phylogenet Evol* 113:33–48.
7. Faircloth BC, Sorenson L, Santini F, Alfaro ME (2013) A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). *PLoS One* 8:e65923.
8. Harrington RC, et al. (2016) Phylogenomic analysis of carangimorph fishes reveals flatfish asymmetry arose in a blink of the evolutionary eye. *BMC Evol Biol* 16:224.
9. Arcila D, et al. (2017) Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. *Nat Ecol Evol* 1:20.
10. Faircloth BC, et al. (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst Biol* 61:717–726.
11. Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol* 61:727–744.
12. Song S, Zhao J, Li C (2017) Species delimitation and phylogenetic reconstruction of the sinipercids (Perciformes: Sinipercidae) based on target enrichment of thousands of nuclear coding sequences. *Mol Phylogenet Evol* 111:44–55.
13. Simion P, et al. (2017) A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. *Curr Biol* 27:958–967.
14. Vandepoele K, De Vos W, Taylor JS, Meyer A, Van de Peer Y (2004) Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc Natl Acad Sci USA* 101: 1638–1643.
15. Dehal P, Boore JL (2005) Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* 3:e314.
16. Macqueen DJ, Johnston IA (2014) A well-constrained estimate for the timing of the salmonid whole genome duplication reveals major decoupling from species diversification. *Proc R Soc B* 281:20132881.
17. Fitch WM (1970) Distinguishing homologous from analogous proteins. *Syst Zool* 19: 99–113.
18. Li C, Ortí G, Zhang G, Lu G (2007) A practical approach to phylogenomics: The phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evol Biol* 7:44.
19. Edwards SV, Cloutier A, Baker AJ (2017) Conserved nonexonic elements: A novel class of marker for phylogenomics. *Syst Biol* 66:1028–1044.
20. Sun Y, et al. (2016) Fish-T1K (Transcriptomes of 1,000 Fishes) Project: Large-scale transcriptome data for fish evolution studies. *Gigascience* 5:18.
21. Malmstrøm M, et al. (2016) Evolution of the immune system influences speciation rates in teleost fishes. *Nat Genet* 48:1204–1210.
22. Koepfli K-P, Paten B, O'Brien SJ; Genome 10K Community of Scientists (2015) The Genome 10K Project: A way forward. *Annu Rev Anim Biosci* 3:57–111.
23. Betancur-R R, Li C, Munroe TA, Ballesteros JA, Ortí G (2013) Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). *Syst Biol* 62:763–785.
24. Giles S, Xu G-H, Near TJ, Friedman M (2017) Early members of 'living fossil' lineage imply later origin of modern ray-finned fishes. *Nature* 549:265–268.
25. Li C, Riethoven JJM, Naylor GJP (2012) EvolMarkers: A database for mining exon and intron markers for evolution, ecology and conservation studies. *Mol Ecol Resour* 12:967–971.
26. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492–508.
27. Salichos L, Rokas A (2013) Inferring ancient divergences requires genes with strong phylogenetic signals. *Nature* 497:327–331.
28. Salichos L, Stamatakis A, Rokas A (2014) Novel information theory-based measures for quantifying incongruence among phylogenetic trees. *Mol Biol Evol* 31:1261–1271.
29. Zhong B, Betancur-R R (2017) Expanded taxonomic sampling coupled with gene genealogy interrogation provides unambiguous resolution for the evolutionary root of angiosperms. *Genome Biol Evol* 9:3154–3161.
30. Shen XX, Hittinger CT, Rokas A (2017) Contentious relationships in phylogenomic studies can be driven by a handful of genes. *Nat Ecol Evol* 1:126.
31. Arratia G (1997) Basal teleosts and teleostean phylogeny. *Palaeo Ichthyol* 7:5–168.
32. Bian C, et al. (2016) The Asian arowana (*Scleropages formosus*) genome provides new insights into the evolution of an early lineage of teleosts. *Sci Rep* 6:24501.
33. Fink SV, Fink WL (1981) Interrelationships of the ostariophysan fishes (Teleostei). *Zool J Linn Soc* 72:297–353.
34. Chakrabarty P, et al. (2017) Phylogenomic systematics of ostariophysan fishes: Ultraconserved elements support the surprising non-monophyly of characiformes. *Syst Biol* 66:881–895.
35. Wainwright PC, et al. (2012) The evolution of pharyngognathy: A phylogenetic and functional appraisal of the pharyngeal jaw key innovation in labroid fishes and beyond. *Syst Biol* 61:1001–1027.
36. Friedman M (2010) Explosive morphological diversification of spiny-finned teleost fishes in the aftermath of the end-Cretaceous extinction. *Proc R Soc B Biol Sci* 277: 1675–1683.
37. Roch S, Warnow T (2015) On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. *Syst Biol* 64:663–676.
38. Philippe H, et al. (2004) Phylogenomics of eukaryotes: Impact of missing data on large alignments. *Mol Biol Evol* 21:1740–1752.
39. Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol Syst* 44:99–121.
40. Ravi V, Venkatesh B (2018) The divergent genomes of teleosts. *Annu Rev Anim Biosci* 6:47–68.
41. Wheeler TJ, Eddy SR (2013) nhmmer: DNA homology search with profile HMMs. *Bioinformatics* 29:2487–2489.
42. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30:1312–1313.
43. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.
44. Mirarab S, et al. (2014) ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.
45. Kozlov AM, Aberer AJ, Stamatakis A (2015) ExaML version 3: A tool for phylogenomic analyses on supercomputers. *Bioinformatics* 31:2577–2579.
46. Aberer AJ, Kobert K, Stamatakis A (2014) ExaBayes: Massively parallel Bayesian tree inference for the whole-genome era. *Mol Biol Evol* 31:2553–2556.
47. Kobert K, Salichos L, Rokas A, Stamatakis A (2016) Computing the internode certainty and related measures from partial gene trees. *Mol Biol Evol* 33:1606–1617.
48. Yang Z (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 24:1586–1591.