doi:10.1111/evo.13649

Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: The monophyly of characiform fishes

Ricardo Betancur-R.,[1,2,3,4,*] Dahiana Arcila,[2,3,5,*] Richard P. Vari,[5,†] Lily C. Hughes,[3,6] Claudio Oliveira,[7] Mark H. Sabaj,[8] and Guillermo Ortí[3,6]

[1]*Department of Biology, University of Puerto Rico, Río Piedras Campus, San Juan, Puerto Rico 00931*

[2]*Department of Biology, University of Oklahoma, Norman, Oklahoma 73019*

[3]*Department of Vertebrate Zoology, National Museum of Natural History Smithsonian Institution, Washington, DC 20013*

[4]*E-mail: betanri@fishphylogeny.org*

[5]*Sam Noble Oklahoma Museum of Natural History, University of Oklahoma, Norman, Oklahoma 73019*

[6]*Department of Biological Sciences, The George Washington University, Washington, DC 20052*

[7]*Departamento de Morfologia, Instituto de Biociências, Universidade Estadual Paulista, Botucatu, Brazil*

[8]*Department of Ichthyology, The Academy of Natural Sciences of Drexel University, Philadelphia, Pennsylvania 19103*

Received April 30, 2018
Accepted November 5, 2018

Phylogenomic studies using genome-wide datasets are quickly becoming the state of the art for systematics and comparative studies, but in many cases, they result in strongly supported incongruent results. The extent to which this conflict is real depends on different sources of error potentially affecting big datasets (assembly, stochastic, and systematic error). Here, we apply a recently developed methodology (GGI or gene genealogy interrogation) and data curation to new and published datasets with more than 1000 exons, 500 ultraconserved element (UCE) loci, and transcriptomic sequences that support incongruent hypotheses. The contentious non-monophyly of the order Characiformes proposed by two studies is shown to be a spurious outcome induced by sample contamination in the transcriptomic dataset and an ambiguous result due to poor taxonomic sampling in the UCE dataset. By exploring the effects of number of taxa and loci used for analysis, we show that the power of GGI to discriminate among competing hypotheses is diminished by limited taxonomic sampling, but not equally sensitive to gene sampling. Taken together, our results reinforce the notion that merely increasing the number of genetic loci for a few representative taxa is not a robust strategy to advance phylogenetic knowledge of recalcitrant groups. We leverage the expanded exon capture dataset generated here for Characiformes (206 species in 23 out of 24 families) to produce a comprehensive phylogeny and a revised classification of the order.

KEY WORDS: **Exon capture, gene genealogy interrogation, Otophysa, UCEs.**

Phylogenomic analysis, currently comparing hundreds or thousands of genes for an increasingly large number of non-model species, has become mainstream to resolve phylogenetic relationships. Large amounts of information contained in the ever-increasing size of datasets hold the potential to overcome stochastic error arising from limited signal in smaller datasets (Song et al. 2012; Salichos and Rokas 2013; Edwards et al. 2015; Whelan et al. 2015; Simmons et al. 2016; Arcila et al. 2017; Shen et al. 2017; Zhong and Betancur-R. 2017). The appeal to collect and analyze large amounts of data also has been driven by new sequencing technologies, increasing confidence among researchers that datasets sporting genome-scale

*These authors contributed equally to this work.

†Deceased (15 January 2016).

information will unambiguously resolve incongruences in molecular phylogenies. But several recent studies show that phylogenomics can lead to strong support for conflicting hypotheses, indicating that incongruence remains a problem in the era of big genomic data (e.g., Ryan et al. 2013; Dunn et al. 2014; Pisani et al. 2015; Whelan et al. 2015; Arcila et al. 2017; Chakrabarty et al. 2017). Although increasing amounts of information certainly reduce the probability of stochastic error, additional sources of error specific to big data have been identified, generating new and unanticipated challenges and pitfalls for phylogenomic analyses (Philippe et al. 2017).

The development of appropriate inference methods for large datasets has lagged behind the pace of data production, and many current methods originally designed to handle smaller datasets do not scale-up for efficient and consistent analyses in phylogenomics. Conflicting results may stem from model misspecifications leading to systematic biases that are amplified in analyses of multilocus concatenated datasets, effectively distorting signal-to-noise ratios and often revealing high support values for incongruent clades (Philippe et al. 2011; Chiari et al. 2012; Arcila et al. 2017; Reddy et al. 2017). Complex models and partitioning schemes that may properly account for the complexity and heterogeneity of large datasets are generally inapplicable due to unsurmountable computational burden. In most cases, discussions of contradictory results have focused on the relative merits of few methodological options, such as comparing concatenated data analyses versus species–tree inference using the multispecies coalescent (Edwards 2009; Song et al. 2012; Edwards et al. 2015; Springer and Gatesy 2015). While coalescent-based approaches that use gene trees as input are theoretically robust to some systematic biases, they rely on accurate gene tree estimation—an unrealistic assumption in the study of ancient divergences given the low information content of short coalescent genes or c-genes (Betancur-R. et al. 2014; Mirarab et al. 2014; Chou et al. 2015; Roch and Warnow 2015; Simmons et al. 2016). Coestimation of gene trees and the species tree (e.g., using *BEAST, Heled and Drummond 2010) is robust to this problem but inapplicable to large datasets.

Additional sources of systematic error have long been known and attributed to factors such as poor taxonomic sampling (e.g., Heath et al. 2008; Branstetter et al. 2017) or other dataset characteristics such as the amount of missing information (e.g., Philippe et al. 2004; Wiens and Morrill 2011; Hosner et al. 2016). The appeal of diminishing stochastic error by sequencing many genes often compromises adequate taxonomic sampling. Genes or fragments analyzed also need to be orthologous, but the criteria and procedures used to assign genes to orthologous groups are highly diverse and inconsistent among studies and may lead to undetected paralogy (e.g., Hughes et al. 2018). Most importantly, the interaction of all these factors in nonintuitive ways may exacerbate the chance of systematic error (e.g., long-branch attraction, nonstationarity), that is, in turn, amplified by the magnitude and complexity of large datasets.

Phylogenomic analyses also may be plagued by "data errors" (Philippe et al. 2011; Philippe et al. 2017). This source of error originates as undetected mistakes during the construction of the datasets due to poorly designed bioinformatic pipelines that do not routinely apply stringent quality-control steps and because manual curation of the data, such as customarily used in single-gene or small multigene datasets, can become intractable (Philippe et al. 2011; Laurin-Lemay et al. 2012). Standard bioinformatic tools adapted to the size and complexity of the genomic scale are still lacking. Common sources of data error involve cross-contamination or mislabeling of samples, contamination with parasitic organisms, and misalignment involving frameshifts for protein-coding genes due to spurious assemblies or sequencing and annotation errors.

These potential sources of error can be diminished or removed, for example, by filtering the original dataset from "outlier loci" or analyzing only subsets of the most "informative" or "stationary" loci (Townsend 2007; Dornburg et al. 2014; Brown and Thomson 2017; Walker et al. 2018). Even so, the treatment of incongruence may benefit from proposals to use explicit hypothesis-testing procedures that are appropriate for phylogenomic datasets. Likelihood-based tests of topologies in phylogenetics have been very popular for over two decades (Shimodaira and Hasegawa 1999; Goldman et al. 2000; Shimodaira and Hasegawa 2001), but the direct application of such testing procedures to genome-scale datasets is rare or of limited value due to systematic biases, as discussed above. New approaches have been proposed recently to quantify phylogenetic signal and to detect genes or sites that might give rise to incongruence in large multilocus datasets (Salichos and Rokas 2013; Salichos et al. 2014; Arcila et al. 2017; Brown and Thomson 2017; Shen et al. 2017; Walker et al. 2018). Among these, Gene Genealogy Interrogation (GGI) applies the well-known approximately unbiased (AU) test (Shimodaira 2002) in a maximum-likelihood framework to identify the genealogical history that each gene supports with highest probability (Arcila et al. 2017; Zhong and Betancur-R. 2017). The GGI approach extracts the signal from individual genes by defining a reduced set of topological constraints representing alternative resolutions around the conflicting node(s). Because many subclades are usually well supported and uncontroversial, except the ones implied by the incongruent hypotheses, the topological constraints reduce the universe of possible solutions to simpler *n*-taxon statements that can be statistically compared. The AU test is used to rank alternative solutions according to their *P*-values and to reject certain hypotheses in favor of others according to the information contained in each gene. With this procedure, the distribution of support for

alternative topologies across the data matrix is explicitly revealed and may be used directly to identify a top-ranking hypothesis.

GGI also can be used to mitigate the effects of gene tree estimation error in coalescent analyses. The highest ranking constrained gene trees from each locus (the "GGI gene trees") may be used as input for summary coalescent analyses such as ASTRAL-II (Mirarab and Warnow 2015) to obtain a "GGI-based species tree" (Mirarab 2017). Another method similar to GGI was independently proposed (named ΔGLS or ΔSLS depending on whether gene- or site-likelihood scores are used) to dissect the phylogenetic signal in large datasets (Shen et al. 2017). The two approaches differ mainly in that GGI considers all possible topologies for a specific *n*-taxon problem, whereas ΔGLS is based on pairwise comparison between two alternative trees. Topology testing seems a promising avenue to settle cases with conflicting results due to gene tree incongruences. In this study, we demonstrate the power and limitations of GGI and data error curation to investigate the source of conflict among phylogenomic analyses using an iconic clade of freshwater fishes that also has been a model for analysis of Southern Hemisphere biogeographic patterns (Lundberg 1993; Orti and Meyer 1997; Sanmartin and Ronquist 2004; Arroyave et al. 2013; Chen et al. 2013).

The order Characiformes or characins (including tetras, piranhas, and hatchetfishes) has become a contentious clade in the Tree of Life, providing an attractive model to explore factors affecting phylogenomic inference. Characiform species are primary freshwater forms, confined to rivers and lakes in Africa and the Neotropics, with little or no tolerance to salt water. They are split into two well-characterized and undisputedly monophyletic suborders: the African Citharinoidei with about 110 species in two families, and the Characoidei with over 2000 species in 22 families, of which only two are endemic to Africa and the remaining are Neotropical. Extensive morphological evidence (Fink and Fink 1981; Fink and Fink 1996) unambiguously supports the monophyly of the order and its phylogenetic placement in the section Otophysa—the largest radiation of freshwater fishes including over 10,000 species—among the orders Cypriniformes (minnows, loaches, and suckers), Gymnotiformes (South American electric fishes), and Siluriformes (catfishes). Molecular evidence based on a single (Orti and Meyer 1997) or a few DNA markers (Nakatani et al. 2011; Chen et al. 2013), however, has failed to support or outright challenged characiform monophyly and the canonical morphological phylogeny by placing the two characiform suborders closer to other otophysan groups (Siluriformes or Gymnotiformes) rather than to each other. Short internodes connecting the four major otophysan lineages involved (Citharinoidei, Characoidei, Gymnotiformes, and Siluriformes) underscore the difficulty in resolving their relationships (Fig. 1).

The molecular data challenging characiform monophyly have recently escalated to genome-wide proportions due to publication of four phylogenomic studies addressing otophysan phylogeny. Two of these (Arcila et al. 2017; Hughes et al. 2018) have proposed that characoids and citharinoids are sister taxa supporting a monophyletic order Characiformes (MC hereafter), whereas the other two (Chakrabarty et al. 2017; Dai et al. 2018; Fig. 1; Table 1) claimed the non-monophyly of the order (NMC hereafter). The properties of markers examined (exons, transcriptomes, and ultraconserved elements or UCEs) and the number of taxa and loci sampled differ substantially among studies (Table 1). Several alternative analyses applied to these datasets (concatenation or coalescent based) result in significant incongruence, clearly revealing high levels of intra-dataset conflict leading to method-dependent results (e.g., see figure 1 in Arcila et al. 2017). Nonetheless, implementation of GGI in two of these studies finds an overwhelming number of gene trees (>70%) supporting the MC hypothesis, in agreement with GGI-based species tree approaches (Arcila et al. 2017; Hughes et al. 2018) and the canonical morphological hypothesis.

A remarkable feature of the two phylogenomic studies proposing the NMC hypothesis is that they support contrasting topologies, with one resolving Siluriformes as the sister group of Characoidei (Chakrabarty et al. 2017) and the other placing Gymnotiformes in this position (Dai et al. 2018). Also, notable is that these studies examined a small number of otophysan taxa (12–28 vs. 32–225 species in the MC studies; Table 1), and one of them reports a suspiciously shallow divergence between *Phenacogrammus* and *Apteronotus*, the sole representatives of characoids and gymnotiforms examined, respectively (Dai et al. 2018; Fig. 1). Although most time-calibrated analyses of ray-finned fishes indicate that interordinal divergences in Otophysa are at least 100 Ma (Near et al. 2012; Betancur-R. et al. 2013; Chen et al. 2013; Hughes et al. 2018), Dai et al. (2018) date the split between the African *Phenacogrammus* (Characiformes) and the South American *Apteronotus* (Gymnotiformes) at just 28 Ma.

Here, we use hypothesis-testing procedures to reconcile conflicting results obtained with alternative phylogenomic datasets that address characiform monophyly. First, we conducted GGI on the UCE dataset (567 loci) supporting the NMC hypothesis (Chakrabarty et al. 2017) to quantify the level of conflict among UCE loci and the evidence for alternative hypotheses. Second, we address the effects of taxonomic and gene sampling on the statistical power of GGI by subsampling and by expanding the original exon-capture dataset of Arcila et al. (2017), with new data generated for 96 additional characiform species. Third, by reexamining the raw transcriptomic data, we also test whether the shallow interordinal divergences inferred by Dai et al. (2018) are due to contamination of sequence data or taxon misidentification. Finally, using the most extensive genetic and taxonomic coverage to date, we infer a new phylogeny for Characiformes and

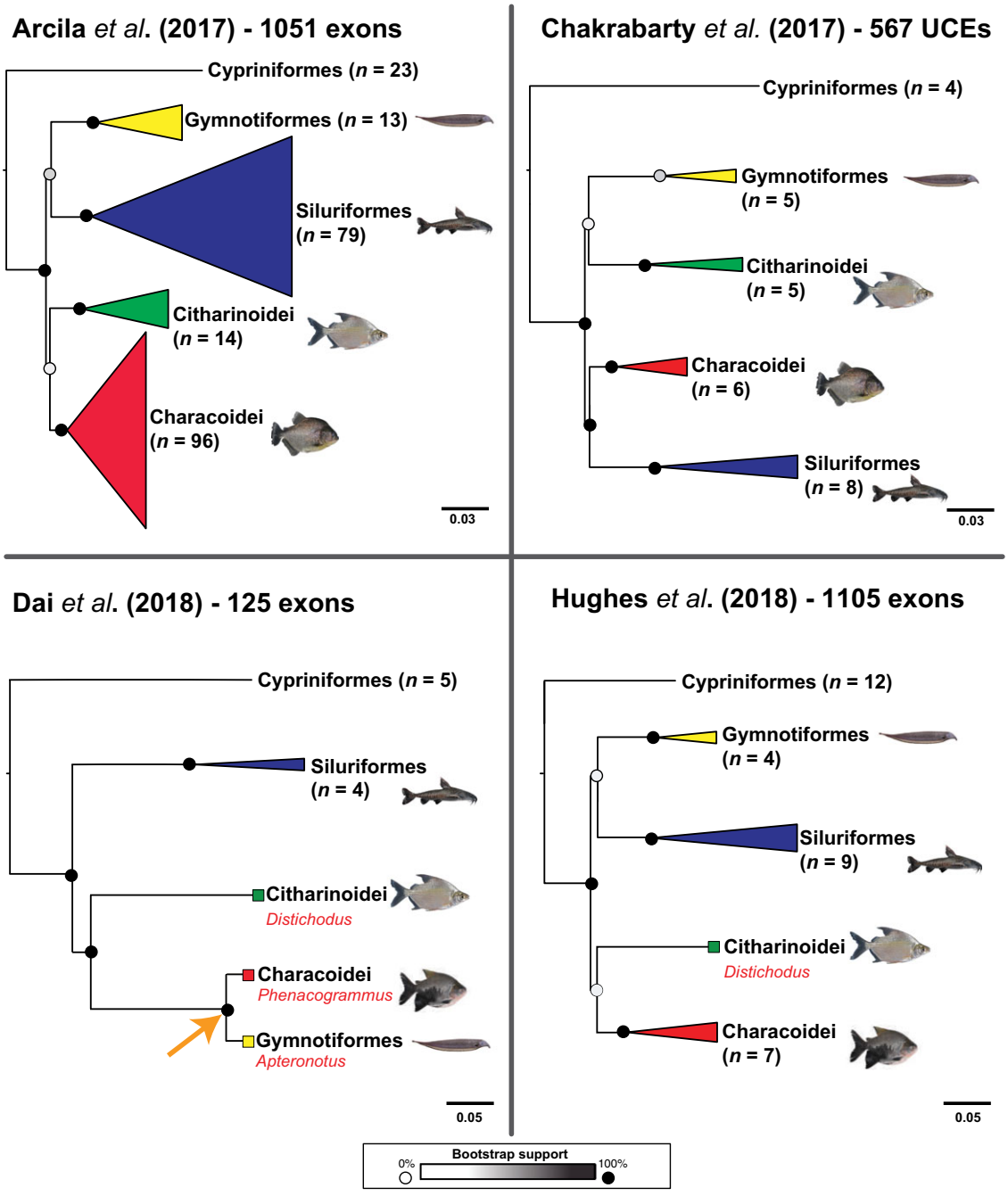

Figure 1. Favored trees inferred by recent phylogenomic analyses of contentious clades of Otophysa. Two studies (Arcila et al. 2017 and Hughes et al. 2018) support the MC hypothesis (monophyletic Characiformes) based on GGI and maximum likelihood (ML) analysis of the concatenated amino acid sequences (also by other results based on species-tree analyses, see figure 1 of Arcila et al. 2017 and figure 2 of Hughes et al. 2018). In contrast, Chakrabarty et al. (2017) and Dai et al. (2018) support the NMC hypothesis (non-monophyletic Characiformes) based on ML and species-tree analyses of DNA sequences (they did not conduct GGI analyses). Note that other non-favored analyses resolve in some cases incongruent trees (e.g., Arcila et al., 2017). Branch support for Arcila et al. (2017) is based on concatenated protein subset that includes only genes with at least 200 species (analysis 21). Orange arrow highlights the suspiciously shallow divergence (short branch lengths) inferred by Dai et al. (2018) between *Phenacogrammus* (Characiformes) and *Apteronotus* (Gymnotiformes). See Table 1 for a summary of phylogenomic dataset properties.

Table 1. Summary of dataset properties and hypotheses from phylogenomic studies that favor (MC) or reject (NMC) the monophyly of Characiformes. A total of 143 exons are common between the dataset marked with the dagger (†) and those marked with the subscript (§).

Study	Species sampled Otophysa	Species sampled Characiformes	Markers	No. of Loci	Alignment length (sites)	Hypothesis supported	Sister group of Characoidei
Arcila et al. (2017)	225	110	Exons	1051[§]	279,012	MC[1]	Citharinoidei
Chakrabarty et al. (2017)	28	11	UCEs	567	145,897	NMC[2]	Siluriformes
Hughes et al. (2018)	32	8	Exons	1105[†]	555,288	MC[3]	Citharinoidei
Dai et al. (2018)	12	2	Exons	125	152,223	NMC[4]	Gymnotiformes
This study	321	206	Exons	1051[§]	279,775	MC[5]	Citharinoidei

[1]MC obtained with GGI (both DNA and proteins). Supported by additional datasets and analyses: analysis 21—subset with at least 200 species (proteins, ML), analyses 25 and 26—complete dataset (DNA, STAR, and NJ-ST, respectively), analysis 29—subset with conserved genes (DNA, ASTRAL-II), analysis 36—complete dataset (proteins, ASTRAL-II), analysis 43—subset with at least 200 species (proteins, ASTRAL-II), and analysis 44—subset with genes common to other studies (proteins, ASTRAL-II).

[2]NMC resolved with the UCE dataset using Bayesian and ML (75% complete, with and without morphology) and ASTRAL-II (dataset 50% and 75% complete).

[3]MC obtained with ML (proteins) and GGI (both DNA and proteins).

[4]NMC obtained using the complete concatenated dataset in BEAST and ML (DNA).

[5]MC resolved with GGI (DNA).

briefly discuss some of the implications of this new hypothesis for biogeographic models and classification.

Methods

TAXONOMIC SAMPLING, EXON CAPTURE AND SEQUENCING

Tissues for 96 species of characoids were obtained from existing collections, most of which have voucher specimens deposited in natural history museums (see Table S1 for a list of material examined). For each sample, genomic DNA was extracted from fin or muscle tissue using a phenol–chloroform protocol in the Autogen© platform. Library preparation, target enrichment, and Illumina sequencing (single-end) was outsourced to Rapid Genomics (http://www.rapid-genomics.com) and followed the same lab protocols and bioinformatic pipelines used in a previous study to obtain sequence data for 1051 exons (Arcila et al. 2017). All alignments were visually inspected (and edited) to check for open reading frames. The 96 species sequenced for this study were combined with a previously published exon-capture dataset (Arcila et al. 2017) to assemble the largest Otophysan dataset for phylogenomic inference to date. This expanded taxonomic sampling consists of 321 individuals representing 318 distinct species-level taxa, of which 206 are characiforms (173 genera), 21 are cypriniforms (21 genera), 13 are gymnotiforms (10 genera), and 78 are siluriforms (77 genera). The order Characiformes is represented by 23 (out of 24) families; the only family not sampled here is the Neotropical and monotypic Tarumaniidae, hypothesized to be the sister group of Erythrinidae (de Pinna et al. 2017; Arcila et al. 2018). Among the newly sequenced data, two species had to be excluded due to cross-contamination detected by analyses of COI sequences (barcodes) obtained with the exon-capture probe set. This quality-control step is routinely used to verify the identity of tissue samples by BLASTing barcodes against public references.

PHYLOGENETIC ANALYSES OF THE EXPANDED DATASET

Maximum likelihood (ML) analyses were conducted in RAxML version 8.2 (Stamatakis 2014) for the concatenated datasets and for each of the 1051 loci to obtain individual genes trees. For the concatenated dataset, RAxML analyses under the GTRGAMMA model (partitioned by gene and codon position) were replicated 30 times and the best scoring tree across searches was selected. Branch support was assessed using the rapid bootstrap algorithm with 300 replicates; the collection of bootstrapped trees was used to draw bipartition frequencies onto the optimal tree. Individual gene trees were inferred using by-codon partitions in RAxML under the GTRGAMMA model, replicated 30 times to find the best scoring tree for each locus.

Multispecies coalescent analyses were conducted in ASTRAL-II version 5.6 (Mirarab and Warnow 2015), using the individual gene trees obtained in RAxML. The ASTRAL-II analyses were conducted following two strategies to define gene trees for input (see below). All datasets and sequence information are deposited in the Dryad Digital Repository (https://datadryad.org//review?doi=doi:10.5061/dryad.vb76b45).

GENE GENEALOGY INTERROGATION

The GGI approach was applied to the UCE dataset of Chakrabarty et al. (2017) as well as to the expanded exon-capture dataset

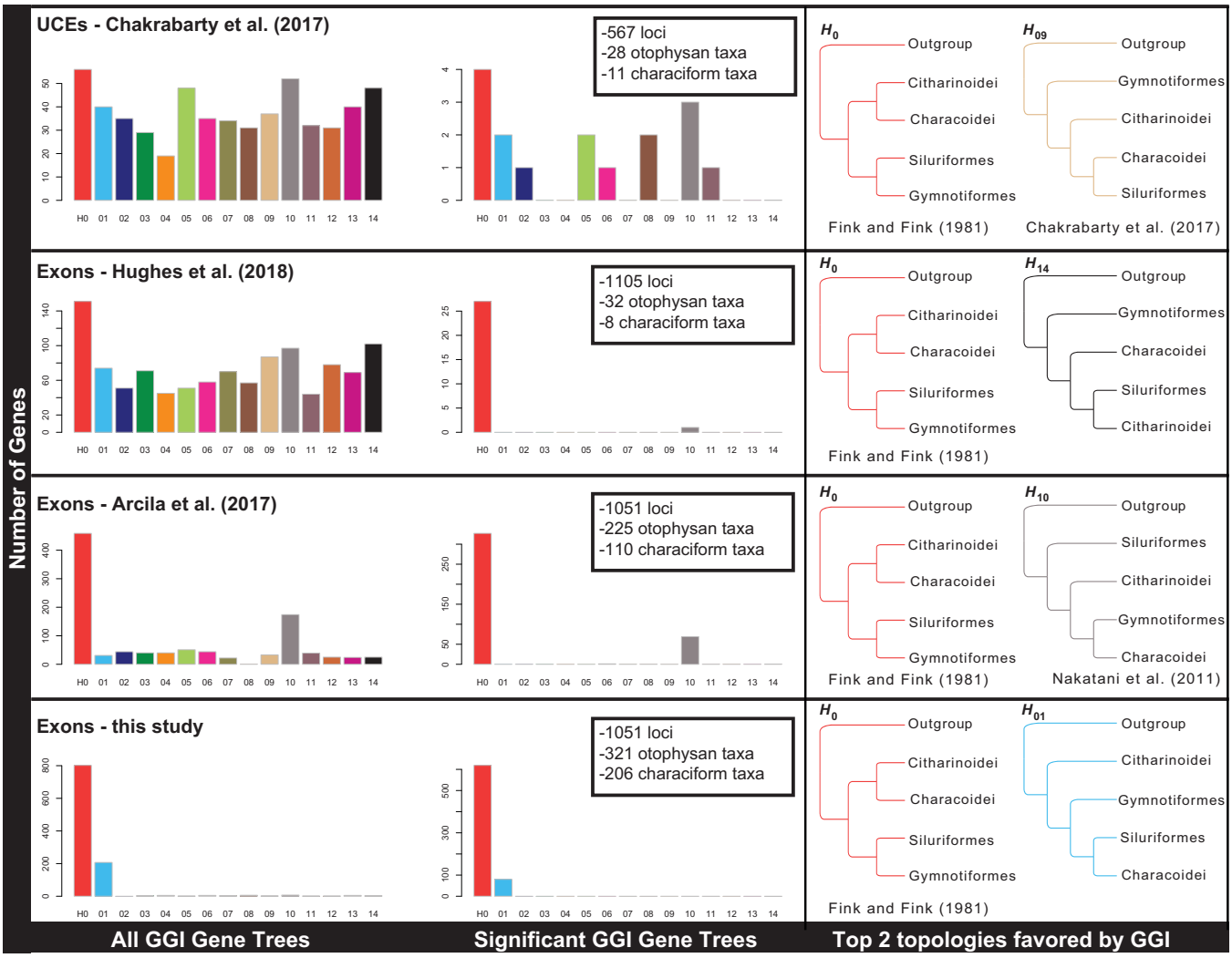

Figure 2. Gene Genealogy Interrogation (GGI) applied to four different datasets. Although differential support in favor of a single tree varies based on different datasets, in all cases the most frequent gene genealogy corresponds to the canonical morphological tree (*H0*), which supports the monophyletic characiforms (MC) hypothesis (tree on the left). The trees on the right are the hypotheses ranked second by GGI test using all GGI gee trees. See Supplementary Table S2 for detailed results on frequencies obtained for *H0*-*H14*.

compiled here and also to a set of replicated subsampled exon datasets (see below). Because GGI analyses were conducted by two other recent studies (Arcila et al. 2017—original exon capture study; Hughes et al. 2018), they are simply reported again here for comparison with the new results (Fig. 2, Table S2). These two analyses used different sets of loci: Arcila et al. (2017) used target capture in combination with exon markers designed specifically for Otophysa, whereas Hughes et al. (2018) compiled exon alignments in silico by mining genome and transcriptome datasets across the fish diversity. Comparing the datasets of these two studies, we find that 143 exons are common to both (Table 1).

For all datasets, the GGI analyses tested the relationships among four lineages (Citharinoidei, Characoidei, Gymnotiformes, and Siluriformes) plus an outgroup (Cypriniformes). These otophysan lineages are uncontroversial and their monophyly is highly supported by different types of data (Fink and Fink 1981; Arcila et al. 2017). They are subtended by long internal branches estimated to represent at least 20–65 million years of evolution, or a conservative estimate of 20–65 coalescent time units (Arcila et al. 2017). This is an important consideration when defining lineages to enforce topological constraints for GGI as the method assumes that these subclades are present in all gene trees. Rosenberg (2003) estimated under a neutral coalescent model that the vast majority of genes in a genome (99.999%) require about eight coalescence time units to achieve monophyly (for either one species or two sister species). Although it is possible that for some genes allelic polymorphisms may be maintained by selection for much longer periods of time resulting in deep coalescences, it

seems unlikely that those isolated instances would introduce systematic bias into the GGI procedure.

Based on these five lineages, we defined rooted four-taxon (or unrooted five-taxon) statements for which there are 15 possible topologies that can be tested using constrained ML searches with RAxML version 8.2 (Stamatakis 2014). Hence, we constrained each of the major subclades to be monophyletic and conform to one of the 15 possible topologies, but imposed no other constraint with regard to relationships within each clade, nor with respect to branch lengths or model parameters. Site likelihood scores for each alternative constrained topology were obtained with RAxML and a topology test was conducted for each gene by statistically comparing the scores of the 15 trees via the AU test (Shimodaira 2002) as implemented in CONSEL version 0.1 (Shimodaira and Hasegawa 2001). Trees were ranked according to the *P*-values and visualized using box plots. Two alternative approaches were applied for scoring the GGI results: (i) one using all rank 1 (best) gene trees and (ii) another using the set of best gene trees that are significantly better ($P < 0.05$) than the alternatives (resulting in a smaller subset of loci). The two sets of GGI trees selected from the previous step were used as input for ASTRAL-II version 5.6 (Mirarab and Warnow 2015).

GGI SENSITIVITY ANALYSES

Because the UCE dataset has a considerably smaller size than the exon datasets (Table 1), we analyzed the effect of taxon and gene sampling on the power of GGI to discriminate among alternative topologies. First, we compiled five smaller exon-capture datasets by subsampling the original data published by Arcila et al. (2017) to mimic the structure of the UCE dataset of Chakrabarty et al. (2017). A total of 28 (out of 225) otophysan taxa (including four cypriniform outgroups and 24 noncypriniform ingroups) and 567 (out of 1051) genes were randomly selected following the number of lineages sampled per major order/suborder in the UCE study (four cypriniforms, eight siluriforms, six characoids, five citharinoids, and five gymnotiforms). Species within each lineage were carefully selected to represent their phylogenetic diversity; hence, some species are inevitably shared across some data subsets. The proportion of unique taxa across the five subsets is 80%. Exon loci were subsampled based on two metrics: (i) taxonomic completeness (exons with sequence data for less than 12 species were excluded) and (ii) information content (to make the exon data most similar to the UCE data). The average number of variable sites per locus in the UCE versus the exon subsets selected is 122.3 and 117.7, respectively. Note, however, that the complete exon dataset with 225 taxa has an average of 143.6 or 162.0 variable sites per locus when all (1051) or the 567 loci with maximum variability were included, respectively. These statistics suggest that (i) as expected, reducing number of taxa also reduces the information content of individual loci, and (ii) UCE loci are on average 14.7% more variable per taxon than exon loci, a factor that we controlled for in our subsampling.

Second, to assess the sensitivity of GGI-based coalescent analyses to the number of loci, we conducted a jackknifing approach to randomly sample the top-ranking GGI gene trees from our expanded exon dataset (with up to 318 species-level taxa, depending on the number of present taxa per gene). The subsampling was done twice based on (i) the complete set of 1051 top-ranking genes trees, and (ii) the subset of GGI gene trees that were significantly better than the alternatives (696 gene trees). Subsets assembled included 5, 10, 20, 30, 100, 200, 300, and 500 gene trees, with up to five nonoverlapping replicates for each (e.g., the complete set of genes trees includes two replicates of 500 genes, but the significantly better set has only one replicate). See details in Table S3.

REANALYSIS OF THE TRANSCRIPTOMIC DATASET OF Dai et al. (2018)

We downloaded the raw reads of *Apteronotus albifrons* (Gymnotiformes) and *Phenacogrammus interruptus* (Characoidei) from the Sequence Read Archive (accession numbers SRX2479408 and SRX2479409). The sequence reads were quality trimmed in Trimmomatic (Bolger et al. 2014) using default parameters, and assembled separately with Trinity version 2.5.1 (Grabherr et al. 2011). For control purposes, we also examined the transcriptome of the characoid *Astyanax mexicanus*, which was sequenced by an independent study (Pasquier et al. 2016). Mitochondrial barcode COI sequences were downloaded for each species from GenBank (KU568756.1 and KU568963.1) and were subsequently queried against the assembled transcriptomes using BLASTN (Chen et al. 2015).

Results

GGI ANALYSIS OF THE UCE AND EXON DATASETS

Following the original GGI study (Arcila et al. 2017), we constrained each of the four major subclades to be monophyletic (characoids, citharinoids, siluriforms, and gymnotiforms; Fig. 1). These constraints were justified because stem lineages of all major otophysan clades span 20–65 coalescent time units (ctu), confidently meeting the theoretical minima of 8 ctu so that >99.99% of the gene genealogies in a genome are monophyletic (Rosenberg 2003). In total, we conducted 8505 constrained ML searches (15 alternative topologies for each of the 567 UCE loci). We report results for the two alternative approaches for sampling GGI trees: (i) the complete dataset including all top-ranked gene trees (567 GGI UCE trees) and (ii) the subset of top-ranked UCE gene trees that are significantly better ($P < 0.05$) than their alternatives (significant-only GGI UCE trees).

The frequencies of GGI gene trees for each alternative hypothesis in the UCE dataset imply relatively similar levels of support for at least four competing topologies (Fig. 2, Table S2). The most frequent gene genealogy (56 gene trees) corresponds to the canonical topology supported by morphology (MC), although it is followed closely by topology H_{10} (52 gene trees), the NMC topology favored by Chakrabarty et al. (2017) (Fig. 2, Table S1). Significant-only GGI gene trees reveal similar results: four versus three gene trees, respectively (Fig. 2, Table S2). Therefore, GGI results obtained with the UCE dataset (Fig. 2) fail to reject with confidence (or overwhelmingly favor) any of the hypotheses. Similarly, coalescent-based analyses conducted in ASTRAL-II (Mirarab and Warnow 2015) using the GGI gene trees as input resolved the NMC tree favored by Chakrabarty et al. (2017) when using as input all 567 GGI gene trees, but the canonical MC tree was obtained when using as input the 16 significant-only GGI genes trees (Fig. S1).

In contrast to the UCE dataset, GGI results from the expanded exon dataset compiled for this study (318 otophysan taxa including 206 characiforms) show overwhelming support for the MC topology, both when all 1051 GGI gene trees are considered (i.e., 802 vs. 207 loci for best alternative) or when significant-only GGI trees are considered (619 vs. 80 loci; Fig 2, Table S2). Interesting differences are noted between the results of this study and the original one by Arcila et al. (2017). Analyses of the expanded exon dataset find stronger support in favor of the MC tree, with 802 genes (76% of all genes) versus 459 genes (43%) in the original study, and identifies in second place a different NMC hypothesis (H_{01}) compared to the original study (H_{10}). The second hypothesis is supported by only 207 genes (a difference of 595 genes or 57% of all genes) in the expanded dataset, whereas the second hypothesis is supported by 174 genes (a difference of 285 genes or 27% of all genes) in the original dataset. The GGI-based coalescent trees obtained with ASTRAL-II support the MC hypothesis (Fig. S2), in agreement with the result obtained by Arcila et al. (2017). The ML tree obtained by RAxML on the concatenated dataset for Characiformes taxa only is shown in Figure 4.

The different results obtained using GGI on the exon and UCE datasets could be due to marker type, dataset size (i.e., an order of magnitude more taxa and nearly twice as many loci were sampled for exons than UCEs; Table 1), or a combination of factors. Although in all datasets the MC gene tree topology under GGI is the most frequent (Fig. 2), the two cases in which GGI finds the smallest differential support in favor of the MC tree are also those with the fewest number of taxa examined (Chakrabarty et al. 2017; Hughes et al. 2018), suggesting that taxonomic sampling affects the statistical power of GGI to resolve alternative relationships. Note, however, that GGI gene tree frequencies alone should not be equated with phylogenetic resolution, as there are theoretical conditions under which the most frequent gene genealogy conflicts with the underlying species tree (i.e., the anomaly zone; see Arcila et al. 2017 for discussion). Finally, consistent with previous studies (e.g., Saitoh et al. 2003; Nakatani et al. 2011; Chen et al. 2013; Arcila et al. 2017; Chakrabarty et al. 2017), non-GGI ML and coalescent-based analyses of the expanded exon dataset continue favoring NMC topologies (Fig. S4).

TAXONOMIC SAMPLING AND GGI

An ideal approach for teasing apart the effect of marker type (UCE or exons) and dataset size (number of taxa and genes sampled) would be to augment the UCE dataset by adding more genes and more taxa, matching the sampling strategies applied in the exon-based studies. Although doing this is beyond the scope of our study, we instead tested the effects of taxonomic sampling on GGI by subsampling the exon-capture dataset (Arcila et al. 2017). The subsets analyzed have same number of genes (567) and otophysan taxa (28) examined with UCEs. We assembled five replicate subsets by including different species (some are common across subsets) while maintaining the number of genes constant. The 567 genes (out of 1051 total) were selected to maximize data for species sampled and to include a comparable number of informative sites per gene (see Methods for details). In all five replicates, the GGI analyses resulted in muddled resolution of alternative tree topologies (Fig. 3) in comparison with full datasets (Fig. 2, Table S2). Contrary to results obtained from the full exon dataset, an NMC topology (H_{03}) is the most frequent tree in subsets 1 and 5 (winning with 102 and 89 GGI trees, respectively), but the alternative hypotheses H_{05} wins in replicates 3 and 4 (with 64 and 65 GGI trees, respectively) and ties for best hypothesis with H_{10} with 62 GGI trees in replicate 2. The MC hypothesis (H_0) never wins, but receives support from 70 GGI trees in replicate 5. In all cases, the winning hypothesis receives support of less than 18% of all loci. The GGI-based coalescent trees obtained with ASTRAL-II for the five subsets represent four different NMC hypotheses (Fig. S3), with H_{05} and H_{03} also being frequently obtained with all GGI trees and only significant GGI trees as input. None of the winning NMC hypotheses corresponds to the preferred tree (H_{10}) reported by Chakrabarty et al (2017).

In summary, these results indicate that limited taxon sampling has a major effect on the statistical power of GGI. Smaller datasets fail to identify clearly a winning hypothesis while GGI results based on the expanded dataset provide clear resolution and increased support for the MC tree. It remains to be seen whether a more taxonomically rich UCE dataset would provide significant support in favor of the NMC hypothesis (using GGI and other methods).

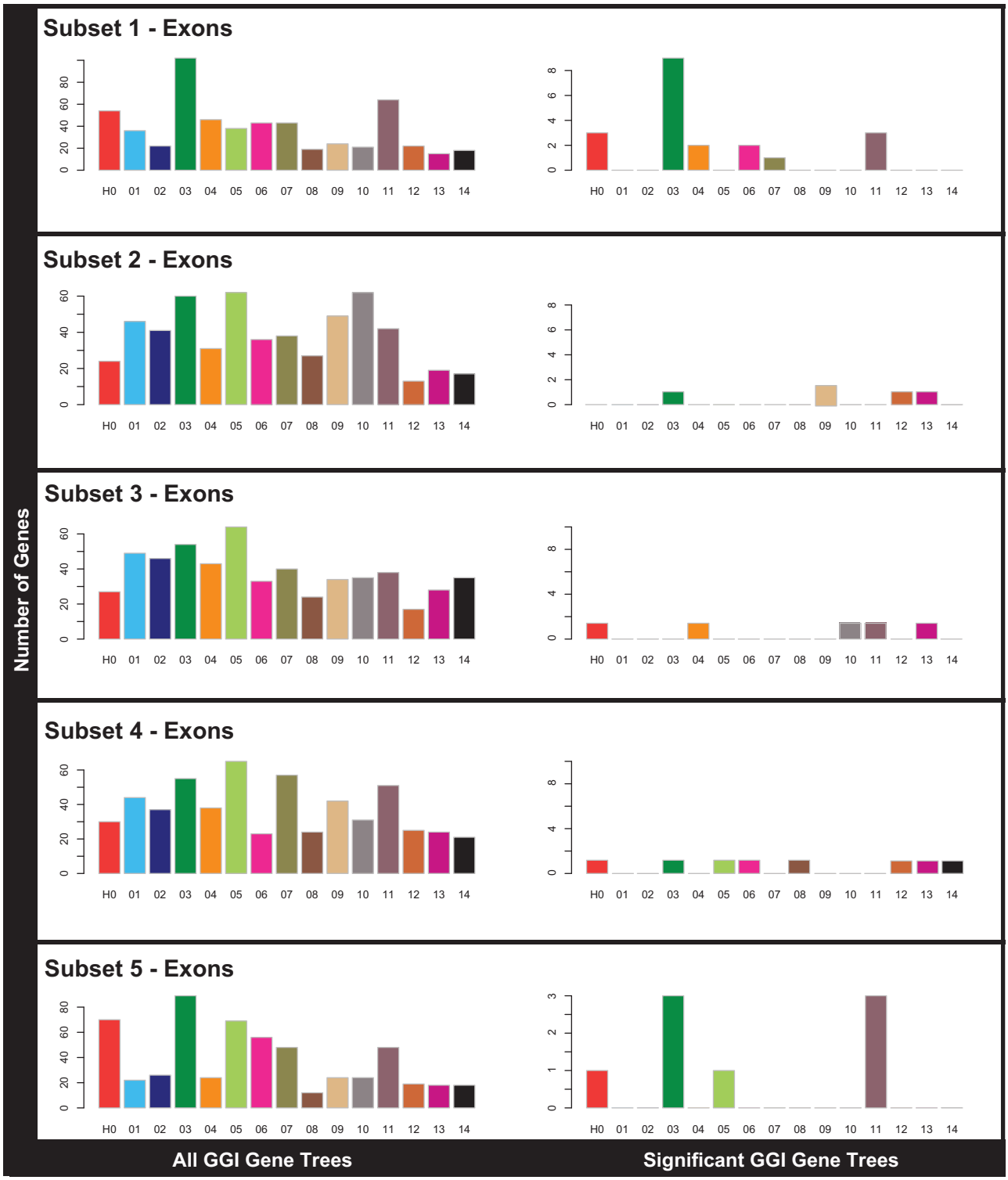

Figure 3. Testing the power of GGI using reduced subsets of the exon capture dataset generated by Arcila et al. (2017). The subsampling targeted 5 sets of 28 otophysan taxa (some are shared across subsets; see Methods) and 567 genes, thereby matching the sampling strategies in the UCE study (Chakrabarty et al. (2017). The relative frequencies of tree topologies supported by the different genes (compared to those using the complete dataset, as well as the expanded dataset generated here; Fig. 2) suggest that GGI is sensitive to limited taxon sampling. See Supplementary Table S1 for details on *H*0-*H*14 and Supplementary Fig. S3 for results of GGI-based coalescent analyses using these subsets.

GENE SAMPLING AND GGI

The effect of varying the number of genes (rather than taxa) on the topology favored by GGI-based coalescent analyses is reported in Table S3. The randomly assembled subsets of gene trees (sampled from the complete set of GGI gene trees) were used as input for ASTRAL-II. Subsets assembled included 5, 10, 20, 30, 100, 200, 300, and 500 gene trees, with up to five replicates each (see Methods for details). For 74 out of 76 GGI gene tree subsets (Table S3), ASTRAL-II supported the MC hypothesis (H_0). Only two species trees obtained with two of the smaller subsets (five and 20 gene trees; Table S3) resolved an NMC tree (H_{03}), which also was the tree favored by ASTRAL-II for the reduced taxon datasets (Fig. S3). In this particular case, these results suggest that decreasing the number of genes on a dataset that otherwise features deep taxonomic coverage has a lesser impact on the GGI-based coalescent results.

CONTAMINATION OF *PHENACOGRAMMUS INTERRUPTUS* IN THE TRANSCRIPTOMIC DATASET

To investigate the unexpectedly shallow divergences between Gymnotiformes and Characoidei in the Dai et al. (2018) study (Fig. 1), we mapped the raw reads of *Apteronotus* (their sole gymnotiform representative) against the assembled transcriptome of *Phenacogrammus* (their sole characoid representative). For control purposes, we also included the transcriptome of another characoid species, *Astyanax mexicanus*, in the comparisons. We also checked for possible instances of cross contamination by querying mitochondrial barcode COI sequences (downloaded from GenBank) against the assembled transcriptomes. In the absence of contamination, we expected to find similar mapping rates of *Apteronotus* reads to both characoid transcriptomes.

The BLASTN query against the *Phenacogrammus* transcriptome detected sequences with 100% identity to both COI barcode sequences, whereas the *Apteronotus* transcriptome only contained the *Apteronotus* COI sequence, suggesting that the *Phenacogrammus* transcriptome is contaminated with *Apteronotus* RNA but not the reverse. Furthermore, a contig assembled in the *Phenacogrammus* transcriptome (>7000 bp) had 100% identity with the *Apteronotus* mitochondrial genome when queried against the NCBI nucleotide database. To further assess the extent of contamination, we quantified the overall alignment rate and found that 86.7% of the raw reads of *Apteronotus* map to the *Phenacogrammus* transcriptome. By contrast, the proportion *Apteronotus* reads mapping against the *Astyanax* transcriptome (characoid control) is only 4.4%.

Collectively, these results indicate that the NMC hypothesis and the shallow divergence between Characoidei and Gymnotiformes (Fig. 1) obtained by Dai et al. (2018) is a spurious result ("data error") arising from massive sequence contamination. The lack of taxonomic replication within each suborder helped conceal this specific instance of contamination. The size and complexity of phylogenomic datasets demand analytical pipelines with stringent quality-control procedures to limit the adverse effects of such data errors (e.g., Philippe et al. 2017).

Discussion

PHYLOGENOMIC INCONGRUENCE AND HYPOTHESIS TESTING

Many factors associated with the construction and analysis of phylogenomic datasets are known to have biasing effects that may lead to conflicting results (Philippe et al. 2011). Most studies, however, typically focus on one or a few of these (if any) in an attempt to increase confidence in the results and advance phylogenetic hypotheses. But negative interactions of identifiable and undetected factors in nonintuitive ways may be amplified by the size and complexity of large datasets, leading to increased reporting of incongruence in the phylogenomic literature. In this study, we address a recent case of incongruence based on analyses of UCE, transcriptome, and exon datasets for the largest clade of freshwater fishes (Otophysa).

We emphasize the importance of hypothesis testing and, in particular, the dissection and characterization of phylogenetic signal contained in large multilocus datasets as a means to determine any internal conflict in each dataset before comparing results from wholesale analysis of different datasets. Our rationale is that systematic biases may override the prevailing signal to produce incongruent results that are not really supported by the different datasets. These systematic biases may take the form of long-branch attraction due to model misspecification, inadequate accounting of base composition nonstationarity, or other factors. By applying a recently developed hypothesis testing procedure (GGI), we show that the apparent incongruence among results reported by Chakrabarty et al. (2017), Arcila et al. (2017), Hughes et al. (2018), and the expanded dataset compiled for this study stems from differences in taxon sampling, rather than quality (UCE or exons) or quantity of loci (500–1000) analyzed. Furthermore, our simple sensitivity analysis shows that the number of taxa included in our dataset has a stronger effect than the number of genes on the power to discriminate among alternative hypotheses using GGI. These results may highlight potential limitations of hypothesis-testing procedures for datasets with small numbers of taxa. Although the original studies supported contradictory results, GGI analyses show that the apparent conflict is spurious. The small dataset (Chakrabarty et al. 2017) simply lacks sufficient information to discriminate among alternative hypotheses despite producing a strongly supported result when analyzed with standard concatenation approaches or species-tree methods such as ASTRAL-II. Furthermore, standard concatenation and

species–tree methods are susceptible to systematic biases and gene tree estimation error, respectively (Meredith et al. 2011; Philippe et al. 2011; Betancur-R. et al. 2014; Springer and Gatesy 2015; Arcila et al. 2017).

In addition to inadequate taxon sampling, another factor that may limit the power of GGI to discriminate among alternative hypotheses is the ability to define unambiguous and well-supported clades to design backbone constraints for the AU tests. There is a compromise between meeting the assumption of long subtending branches (more than eight coalescent time units) at the base of these clades and the number of clades that need to be defined around the node (or nodes) involved in the conflict. If the number of clades that meet this assumption is greater than five (as applied in this study), the number of alternative topologies may become prohibitively high to efficiently apply GGI via AU tests. For five-taxon (or unrooted six-taxon) statements, there are 105 possible topologies, and for six-taxon (or unrooted seven-taxon) statements, there are 945 that would need to be tested for each gene tree in the dataset. Therefore, the ability to "dissect" a subset of well supported clades around the node of interest that meet the monophyly criterion may become another critical limitation of the GGI approach. For example, this seems to be the case with the difficulty of resolving controversial branching patterns proposed by different studies at the base of the neoteleost and acanthopterygian fish radiations (Near et al. 2012; Betancur-R. et al. 2013; Grande et al. 2013; Davesne et al. 2016; Nelson et al. 2016; Alfaro et al. 2018). Fortunately, this is not the case for the apparent incongruence in otophysan relationships or characiform monophyly.

THE IMPORTANCE OF TAXONOMIC SAMPLING

Our confidence in phylogenetic inference clearly increases with the size of the datasets analyzed. Large amounts of information reduce the probability of stochastic error in general, but this study shows that investing resources into maximizing taxonomic sampling may be as relevant as examining genome-scale datasets. We demonstrate that the power to discriminate among competing hypotheses with GGI seems to be more contingent on having a dense taxonomic sample than on the number or type of genes analyzed. Although filtering for informative loci may be a promising avenue to gain confidence in resulting phylogenies by increasing measures of support (bootstrap or other), efforts aimed at including dense taxon coverage are critical when applying GGI to addressing recalcitrant groups in the Tree of Life.

Our analysis of smaller subsets of taxa (Fig. 3) shows that even with the same amount of sequence data (1051 loci in this case), the ability to discriminate among competing hypotheses using topology tests on gene trees (GGI) is lost. It seems that, in spite of retaining taxa for all relevant groups involved in the controversial topological patterns (orders or suborders of otophysan fishes), the AU tests fail to resolve a predominant branching order. A possible explanation is that, just by pure chance, ML scores for alternative trees with a low number of tips will likely be more similar to each other than alternative trees with a large number of tips. In fact, AU tests should be used with caution when many of the best trees are nearly equally as good, as one might overlook the "true tree" by placing too much confidence in the wrong trees (Shimodaira 2002). This problem will affect each individual AU test for all loci (either UCEs or exons) that are short and uninformative or taxonomically undersampled. Our analyses provide support for the latter. We did not explicitly test for locus length or information content, although we did match the information content of UCE and exon loci in the subsampled datasets. Simulation studies under different conditions of taxon sampling and locus informativeness would help tease apart their effects on the power of GGI to discriminate among apparently conflicting phylogenomic hypotheses.

ADDRESSING DATA ERROR

Although not pervasive, a worrisome trend in the phylogenomic literature is the relatively high incidence of data error reported *a posteriori* of the publication of landmark papers. For example, Philippe et al. (2017) discuss three datasets that addressed early metazoan evolution and the origin of land plants in which undetected data errors may have compromised the published results. The errors were carefully uncovered by time-consuming reanalyses conducted by independent authors (Philippe et al. 2011; Laurin-Lemay et al. 2012). Mirarab and Warnow (2015) and Springer and Gatesy (2015) characterized fundamental errors in Song's et al. (2012) phylogenomic dataset (447 genes) that compromised their major conclusions about relationships among eutherian mammals. Additional examples involve the position of turtles in relation to lepidosaurs and archosaurs proposed by analysis of a large microRNA dataset (Lyson et al. 2012) in which the implementation of more rigorous criteria for microRNA annotation and orthology assessment has challenged the original results (Field et al. 2014). Such examples can lead to the publication of "phylogenomic errata" that attempt to correct the original mistakes (e.g., Song et al. 2015), but ultimately add confusion to the Tree of Life.

The challenge to develop robust bioinformatic pipelines to assemble phylogenomic datasets that incorporate quality-control steps at several levels has not been met in practice. Because visual inspection of alignments and reading frames is more difficult for a large dataset with hundreds or thousands of genes, efficient tools that at least flag potential problems in subsets of the data are desirable. Smaller subsets of flagged alignments may be susceptible to manual curation, but not complete datasets. Problems with contamination or confused species labels also need to be addressed

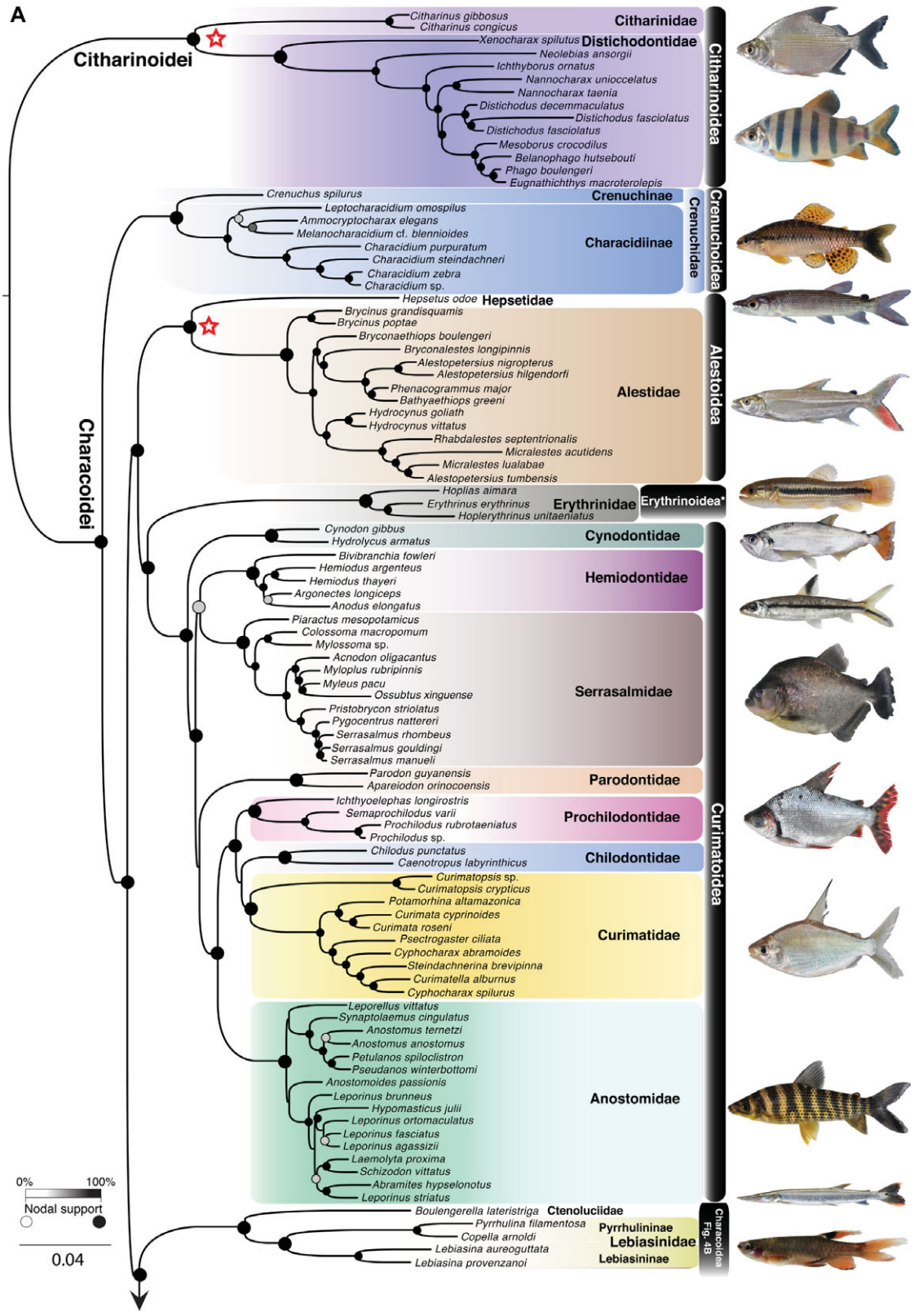

Figure 4. (A) Characiform phylogeny based on the concatenation analysis of 1051 exons. See Supplementary Information for details on the phylogeny and classification of the group. Star highlights African clades; all other groups are Neotropical. (B) Continued on next page. See Supplementary Information for details on the phylogeny and classification of the group. *Erythrinoidea also includes Tarumaniidae (Arcila et al. 2018; see Supplementary Information).

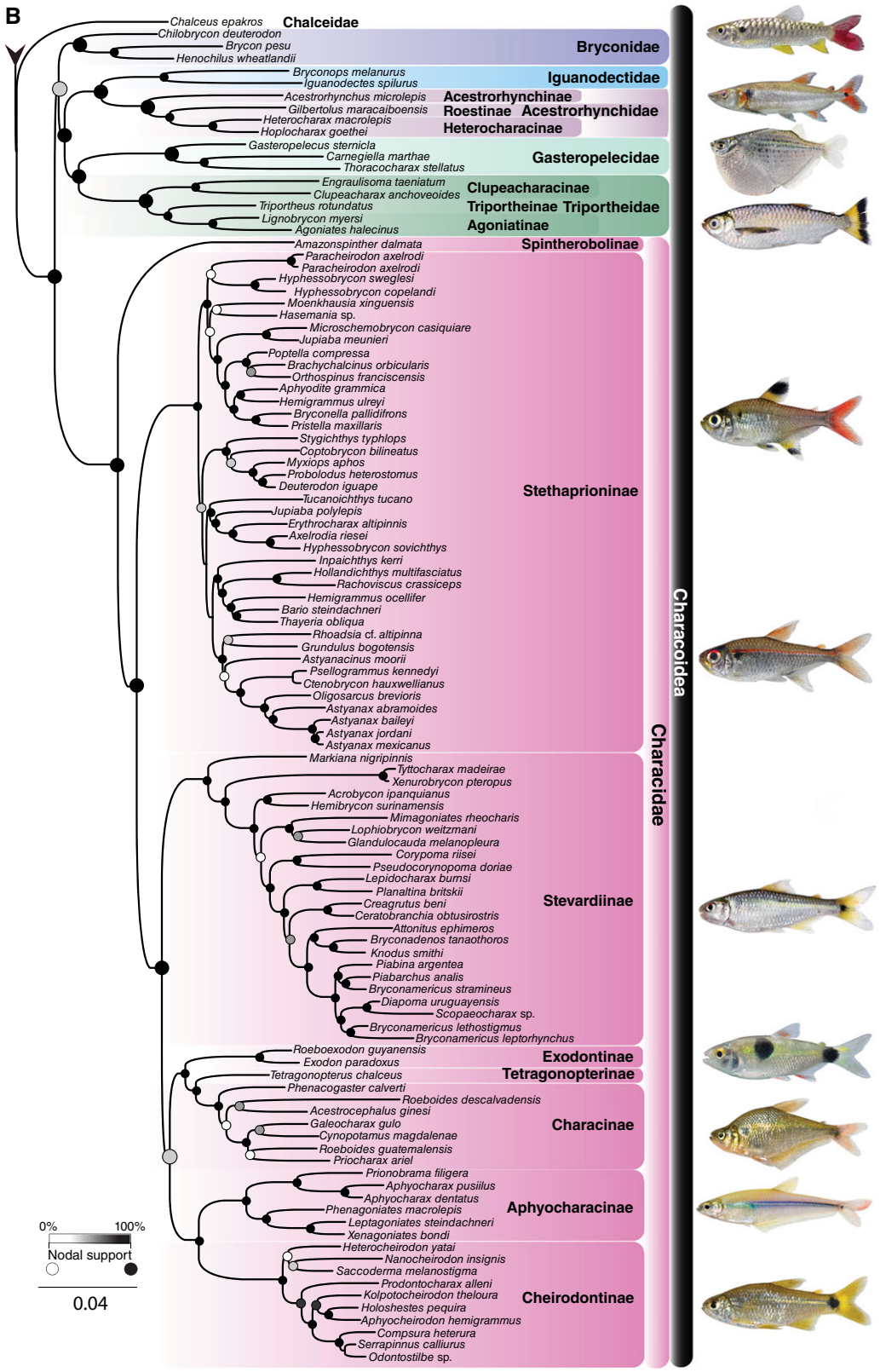

Figure 4. Continued.

by routinely obtaining barcode sequences (typically the mitochondrial COI fragment) that can be quickly compared against curated reference datasets. The simple steps used in this study to scrutinize suspicious results published by Dai et al. (2018) and those used by other authors that reanalyze existing datasets need to be performed routinely and constitute a set of best practice principles for phylogenomics.

CHARACIFORM PHYLOGENY

This study leverages unprecedented amounts of evidence from 1051 exons, generated for 206 characiform species (in 23 out of 24 families), to produce the most comprehensive phylogeny of the order to date (Fig. 4). Our new dataset, along with reanalyses of available UCE data that challenged the monophyly of the order, provide unambiguous support for a long-standing hypothesis based on morphology (Fink and Fink 1981) supporting the monophyly of characiforms and their interrelations among otophysan orders. Our phylogenomic analysis also resolves with confidence the branching pattern among families—particularly toward the base of the characiform radiation—that remained elusive in previous studies (Oliveira et al. 2011), clarifying the relationships among African and Neotropical groups. These results provide a robust phylogenetic framework with dense taxonomic sampling for future comparative studies to address the diversification of characiform fishes. We refer the readership interested in characiform systematics to the Supplementary Information for a revised classification of this group.

Characiforms are among a few classic biogeographic models for the study of broad transoceanic distributions (Sanmartin and Ronquist 2004) because they are primary freshwater forms and cannot be easily transported across the ocean by rafting, wings, or other agents. They have been studied in relation to the biogeographic history of the Southern Hemisphere as a prime example of the vicariance paradigm: disjunct trans-oceanic distributions are a consequence of the sequential breakup of the southern supercontinent of Gondwana. Cretaceous breakup of Gondwana has been proposed as an important mechanism in the diversification of characiform families, but early phylogenies of the group—despite their inability to confidently resolve relationships among all families—suggested that African and Neotropical taxa did not form reciprocally monophyletic groups (Orti and Meyer 1997; Oliveira et al. 2011). These results prompted some authors to identify problems with the vicariance paradigm and to suggest that the present biogeographic distribution implies a highly uneven rate of extinction among African lineages (Lundberg 1993). Chen et al. (2013) found evidence that the diversification of Characoidei postdated the final fragmentation of Gondwana and suggested a preeminent role for uncertain postfragmentation dispersal routes between Africa and South America. Our study confirms to some extent these early suggestions by placing the suborder Citharinoidei, composed of two families endemic to Africa (Citharinidae and Distichodontidae), as the sister group of Characoidei, clade including two African families (Alestidae and Hepsetidae) that are nested within the remaining characiforms, all of which are Neotropical (Fig. 4). The implications of our phylogenetic hypothesis for characiform biogeography, however, need to be elucidated using modern biogeographic methods, a goal that is beyond the scope of this study.

Conclusions

Phylogenomic data have contributed greatly to resolve the structure of the Tree of Life, but several branches remain poorly resolved or subtending contentious relationships. The increased information and power of large datasets invites new challenges to properly curate and analyze them. Efficient analytical approaches that scale-up to large data and, at the same time, accommodate their complexity and heterogeneity remain to be established. The example provided in this article illustrates how hypothesis-testing procedures may be used to reveal the degree of internal conflict (gene-tree discordance) and to resolve conflicting relationships with phylogenomic data. Our results highlight the importance of dense taxon sampling to resolve difficult phylogenetic questions where a "more genes" approach by itself clearly is not sufficient. A well-balanced effort to collect genome-wide datasets for many taxa remains a fundamental step for advancing the field of systematics and, more generally, for unraveling the Tree of Life.

AUTHOR CONTRIBUTIONS

R.B.R. and D.A. contributed equally to this study. R.B.R., D.A., R.V., and G.O. planned and oversaw the project; R.B.R. and D.A. carried out the experiments and collected the data; R.B.R. and D.A. analyzed data; L.H. reanalyzed the Dai et al. dataset. D.A., R.B.R, M.H.S., G.O., and C.O. proposed the revised classification. M.H.S. and C.O. collected, identified, and curated the fish materials examined. R.B.R., D.A., M.H.S., and G.O. wrote the paper and all other authors contributed to the writing.

ACKNOWLEDGMENTS

We thank J. W. Armbruster (Auburn U.) for providing tissue material for five characiform species. We are grateful with J. M. Mirande (Fundación Miguel Lillo) for comments on the revised classification. We thank Jose Carlos Bonilla and Humberto Ortiz for providing support with bioinformatic analyses and access to the High Performance Computing facility of UPR-RP (funded by INBRE Grant Number P20GM103475 from the National Institute of General Medical Sciences and the National Institutes of Health). We also thank members of the RB and GO labs for providing valuable comments on earlier versions of the paper. This project was supported by the National Science Foundation (NSF) grants to R.B.R. (DEB-147184 and DEB-1541491), D.A (DBI-1811748), M.H.S (DEB-1257813) and to G.O. (DEB-1457426 and DEB-1541554). Additional funding was provided by the Opportunity Research Program between George Washington University and the Smithsonian National Museum of Natural History (G.O. and R.V.), the Smithsonian Peter Buck fellowship (R.B.R.), the National Science and Technology Council of the Brazilian

Federal Government (CNPq grant no. 306054/2006-0 to C.O.) and São Paulo State Foundation (FAPESP grant no. 2014/26508-3 to C.O.).

DATA ARCHIVING

The doi for our data is https://doi.org/10.5061/dryad.vb76b45.vb76b45

REFERENCES

Alfaro, M. E., B. C. Faircloth, R. C. Harrington, L. Sorenson, M. Friedman, C. E. Thacker, C. H. Oliveros, D. Černý, and T. J. Near. 2018. Explosive diversification of marine fishes at the Cretaceous–Palaeogene boundary. Nat. Ecol. Evol. 2:688–696.

Arcila, D., G. Ortí, R. Vari, J. W. Armbruster, M. L. J. Stiassny, K. D. Ko, M. H. Sabaj, J. Lundberg, L. J. Revell, and R. Betancur-R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat. Ecol. Evol. 1:0020.

Arcila, D., P. Petry, and G. Orti. 2018. Phylogenetic relationships of the family Tarumaniidae (Characiformes) based on nuclear and mitochondrial data. Neotrop. Ichthyol. 16:180016.

Arroyave, J., J. S. S. Denton, and M. L. J. Stiassny. 2013. Are characiform fishes Gondwanan in origin? Insights from a time-scaled molecular phylogeny of the Citharinoidei (Ostariophysi: Characiformes). PLoS One 8:e77269.

Betancur-R., R., R. E. Broughton, E. O. Wiley, K. Carpenter, J. A. Lopez, C. Li, N. I. Holcroft, D. Arcila, M. Sanciangco, J. Cureton, et al. 2013. The tree of life and a new classification of bony fishes. PLoS Curr. 5. https://doi.org/10.1371/currents.tol.53ba26640df0ccaee75bb165c8c26288

Betancur-R., R., G. Naylor, and G. Orti. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst. Biol. 63:257–262.

Bolger, A. M., M. Lohse, and B. Usadel. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Branstetter, M. G., B. N. Danforth, J. P. Pitts, B. C. Faircloth, P. S. Ward, M. L. Buffington, M. W. Gates, R. R. Kula, and S. G. Brady. 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr. Biol. 27:1019–1025.

Brown, J. M. and R. C. Thomson. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

Chakrabarty, P., B. C. Faircloth, F. Alda, W. B. Ludt, C. D. McMahan, T. J. Near, A. Dornburg, J. S. Albert, J. Arroyave, M. L. J. Styasny, et al. 2017. Phylogenomic systematics of Ostariophysan fishes: ultraconserved elements support the surprising non-monophyly of characiformes. Syst. Biol. 66:881–895.

Chen, W. J., S. Lavoué, and R. L. Mayden. 2013. Evolutionary origin and early biogeography of otophysan fishes (Ostariophysi: Teleostei). Evolution 67:2218–2239.

Chen, Y., W. Ye, Y. Zhang, and Y. Xu. 2015. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 43:7762–7768.

Chiari, Y., V. Cahais, N. Galtier, and F. Delsuc. 2012. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 10:65.

Chou, J., A. Gupta, S. Yaduvanshi, R. Davidson, M. Nute, S. Mirarab, and T. Warnow. 2015. A comparative study of SVDquartets and other coalescent-based species tree estimation methods. BMC Genom. 16: 1–11.

Dai, W., M. Zou, L. Yang, K. Du, W. Chen, Y. Shen, R. L. Mayden, and S. He. 2018. Phylogenomic perspective on the relationships and evolutionary history of the major otocephalan lineages. Sci. Rep. 8:205.

Davesne, D., C. Gallut, V. Barriel, P. Janvier, G. Lecointre, and O. Otero. 2016. The phylogenetic intrarelationships of spiny-rayed fishes (Acanthomorpha, Teleostei, Actinopterygii): fossil taxa increase the congruence of morphology with molecular data. Front. Ecol. Evol. 4:129.

de Pinna, M., J. Zuanon, L. Rapp Py-Daniel, and P. Petry. 2017. A new family of neotropical freshwater fishes from deep fossorial Amazonian habitat, with a reappraisal of morphological characiform phylogeny (Teleostei: Ostariophysi). Zool. J. Linn. Soc. 182:76–106.

Dornburg, A., J. P. Townsend, M. Friedman, and T. J. Near. 2014. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. BMC Evol. Biol. 14:169.

Dunn, C. W., G. Giribet, G. D. Edgecombe, and A. Hejnol. 2014. Animal phylogeny and its evolutionary implications. Ann. Rev. Ecol. Evol. Syst. 45:371–395.

Edwards, S. V. 2009. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Edwards, S. V., Z. Xi, A. Janke, B. C. Faircloth, J. E. McCormack, T. C. Glenn, B. Zhong, S. Wu, E. M. Lemmon, A. R. Lemmon, et al. 2015. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol. Phylogenet. Evol. 94(Pt A): 447–462.

Field, D. J., J. A. Gauthier, B. L. King, D. Pisani, T. R. Lyson, and K. J. Peterson. 2014. Toward consilience in reptile phylogeny: miRNAs support an archosaur, not lepidosaur, affinity for turtles. Evol. Dev. 16:189–196.

Fink, S. V. and W. L. Fink. 1981. Interrelationships of the Ostariophysan Fishes (Teleostei). Zool. J. Linn. Soc. 72:297–353.

Fink, S. V., and W. L. Fink. 1996. Interrelationships of ostariophysan fishes (Teleostei). Pp. 209–249 *in* M. L. J. Stiassny, L. R. Parenti, and G. D. Johnson, eds. Interrelationships of fishes. Academic Press, San Diego, CA.

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.

Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29:644–652.

Grande, T., W. C. Borden, and W. L. Smith. 2013. Limits and relationships of Paracanthopterygii: A molecular framework for evaluating past morphological hypotheses. Pp. 385–418 *in* G. Arratia, H.-P. Schultze, and M. V. H. Wilson, eds. Mesozoic Fishes 5 – global diversity and evolution. Verlag Dr. F. Pfeil, Munich, Germany.

Heath, T. A., S. M. Hedtke, and D. M. Hillis. 2008. Taxon sampling and the accuracy of phylogenetic analyses. J. Syst. Evol. 46:239–257.

Heled, J., and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hosner, P. A., B. C. Faircloth, T. C. Glenn, E. L. Braun, and R. T. Kimball. 2016. Avoiding missing data biases in phylogenomic inference: an empirical study in the landfowl (Aves: Galliformes). Mol. Biol. Evol. 33:1110–1125.

Hughes, L. C., G. Ortí, Y. Huang, Y. Sun, C. C. Baldwin, A. W. Thompson, D. Arcila, R. Betancur-R, C. Li, L. Becker, et al. 2018. Comprehensive phylogeny of fishes (Actinopterygii) based on genomic and transcriptomic data. Proc. Natl. Acad. Sci. USA 115:6249–6254.

Laurin-Lemay, S., H. Brinkmann, and H. Philippe. 2012. Origin of land plants revisited in the light of sequence contamination and missing data. Curr. Biol. 22:R593–R594.

Lundberg, J. G. 1993. African-South American freshwater fish clades and continental drift: problems with a paradigm. Pp. 156–199 *in* P. Goldblatt, ed. Biological relationships between Africa and South America. Yale University Press, New Haven, CT & London, U. K.

Lyson, T. R., E. A. Sperling, A. M. Heimberg, J. A. Gauthier, B. L. King, and K. J. Peterson. 2012. MicroRNAs support a turtle + lizard clade. Biol. Lett. 8:104–107.

Meredith, R. W., J. E. Janečka, J. Gatesy, O. A. Ryder, C. A. Fisher, E. C. Teeling, A. Goodbla, E. Eizirik, T. L. L. Simão, T. Stadler, et al. 2011. Impacts of the cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science 334:521–524.

Mirarab, S. 2017. Phylogenomics: constrained gene tree inference. Nat. Ecol. Evol. 1:56.

Mirarab, S., M. S. Bayzid, B. Boussau, and T. Warnow. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346:1250463.

Mirarab, S., and T. Warnow. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31:i44–i52.

Nakatani, M., M. Miya, K. Mabuchi, K. Saitoh, and M. Nishida. 2011. Evolutionary history of Otophysi (Teleostei), a major clade of the modern freshwater fishes: Pangaean origin and Mesozoic radiation. BMC Evol. Biol. 11:177.

Near, T. J., R. I. Eytan, A. Dornburg, K. L. Kuhn, J. A. Moore, M. P. Davis, P. C. Wainwright, M. Friedman, and W. L. Smith. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc. Natl. Acad. Sci. USA 109:13698–13703.

Nelson, J. S., T. Grande, and M. V. H. Wilson. 2016. Fishes of the world. John Wiley & Sons, Hoboken, NJ.

Oliveira, C., G. S. Avelino, K. T. Abe, T. C. Mariguela, R. C. Benine, G. Orti, R. P. Vari, and R. M. Correa e Castro. 2011. Phylogenetic relationships within the speciose family Characidae (Teleostei: Ostariophysi: Characiformes) based on multilocus analysis and extensive ingroup sampling. BMC Evol. Biol. 11:275.

Orti, G., and A. Meyer. 1997. The radiation of characiform fishes and the limits of resolution of mitochondrial ribosomal DNA sequences. Syst. Biol. 46:75–100.

Pasquier, J., C. Cabau, T. Nguyen, E. Jouanno, D. Severac, I. Braasch, L. Journot, P. Pontarotti, C. Klopp, J. H. Postlethwait, et al. 2016. Gene evolution and gene expression after whole genome duplication in fish: the PhyloFish database. BMC Genom. 17:368.

Philippe, H., H. Brinkmann, D. V. Lavrov, D. T. J. Littlewood, M. Manuel, G. Worheide, and D. Baurain. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9: e1000602.

Philippe, H., E. Snell, E. Bapteste, P. Lopez, P. Holland, and D. Casane. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.

Philippe, H., D. M. d. Vienne, V. Ranwez, B. Roure, D. Baurain, and F. Delsuc. 2017. Pitfalls in supermatrix phylogenomics. Eur. J. Taxon. 283:1–25.

Pisani, D., W. Pett, M. Dohrmann, R. Feuda, O. Rota-Stabelli, H. Philippe, N. Lartillot, and G. Worheide. 2015. Genomic data do not support comb jellies as the sister group to all other animals. Proc. Natl. Acad. Sci. USA 112:15402–15407.

Reddy, S., R. T. Kimball, A. Pandey, P. A. Hosner, M. J. Braun, S. J. Hackett, K. L. Han, J. Harshman, C. J. Huddleston, S. Kingston, et al. 2017. Why do phylogenomic data sets yield conflicting trees? Data type influences the avian tree of life more than taxon sampling. Syst. Biol. 66:857–879.

Roch, S., and T. Warnow. 2015. On the robustness to gene tree estimation error (or lack thereof) of coalescent-based species tree methods. Syst. Biol. 64:663–676.

Rosenberg, N. A. 2003. The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57:1465–1477.

Ryan, J. F., K. Pang, C. E. Schnitzler, A. D. Nguyen, R. T. Moreland, D. K. Simmons, B. J. Koch, W. R. Francis, P. Havlak, N. C. S. Program, et al. 2013. The genome of the ctenophore *Mnemiopsis leidyi* and its implications for cell type evolution. Science 342:1242592.

Saitoh, K., M. Miya, J. G. Inoue, N. B. Ishiguro, and M. Nishida. 2003. Mitochondrial genomics of ostariophysan fishes: perspectives on phylogeny and biogeography. J. Mol. Evol. 56:464–472.

Salichos, L., and A. Rokas. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.

Salichos, L., A. Stamatakis, and A. Rokas. 2014. Novel information theory-based measures for quantifying incongruence among phylogenetic trees. Mol. Biol. Evol. 31:1261–1271.

Sanmartin, I., and F. Ronquist. 2004. Southern Hemisphere biogeography inferred by event-based models: plant versus animal patterns. Syst. Biol. 53:216–243.

Shen, X.-X., C. T. Hittinger, and A. Rokas. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1:0126.

Shimodaira, H. 2002. An approximately unbiased test of phylogenetic tree selection. Syst. Biol. 51:492–508.

Shimodaira, H., and M. Hasegawa. 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114–1116.

Shimodaira, H., and M. Hasegawa. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics 17:1246–1247.

Simmons, M. P., D. B. Sloan, and J. Gatesy. 2016. The effects of subsampling gene trees on coalescent methods applied to ancient divergences. Mol. Phylogenetics Evol. 97:76–89.

Song, S., L. Liu, S. V. Edwards, and S. Wu. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA 109:14942–14947.

Song, S., L. Liu, S. V. Edwards, and S. Wu. 2015. Correction for Song et al., resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proc. Natl. Acad. Sci. USA 112:E6079.

Springer, M. S., and J. Gatesy. 2015. The gene tree delusion. Mol. Phylogenetics Evol. 94:1–33.

Townsend, J. P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56:222–231.

Walker, J. F., J. W. Brown, and S. A. Smith. 2018. Analyzing contentious relationships and outlier genes in phylogenomics. Syst. Biol. 67: 916–924.

Whelan, N. V., K. M. Kocot, L. L. Moroz, and K. M. Halanych. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl. Acad. Sci. USA 112:5773–5778.

Wiens, J. J., and M. C. Morrill. 2011. Missing data in phylogenetic analysis: reconciling results from simulations and empirical data. Syst. Biol. 60:719–731.

Zhong, B., and R. Betancur-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.

Associate Editor: I. Sanmartín
Handling Editor: P. Tiffin

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1. Species-tree analyses using GGI trees conducted in ASTRAL-II using the (Chakrabarty et al. 2017) dataset as input data.
Figure S2. Species-tree analyses using GGI trees conducted in ASTRAL-II for the expanded characiforms exon dataset including all GGI gene trees.
Figure S3. Species-tree analyses using GGI trees conducted in ASTRAL-II for the five exon subsets (567 genes and 28 otophysan taxa).
Figure S4. Otophysan phylogenies of the expanded dataset (318 species), including characiforms and non-characiform outgroups.
Table S1. List of specimens examined and catalog numbers. Museum acronyms are given as footnote.
Table S2. Summary of the results using gene genealogy interrogation (GGI) on the four complete datasets analyzed.
Table S3. GGI-based species tree analyses (ASTRAL-II) obtained with the exon capture dataset that includes all taxa examined (up to 321, depending on the number of present taxa per gene) with varying number of GGI genes (randomly subsampled from the complete set of GGI gene trees).
Table S4. Higher-level classification of Order Characiformes proposed by this study compared to summary of traditional classification.