Resolving Deep Nodes in an Ancient Radiation of Neotropical Fishes in the Presence of Conflicting Signals from Incomplete Lineage Sorting

Fernando Alda[1,2*], Victor A. Tagliacollo[3], Maxwell J. Bernt[4], Brandon T. Waltz[4], William B. Ludt[1], Brant C. Faircloth[1], Michael E. Alfaro[5], James S. Albert[4], Prosanta Chakrabarty[1]

[1]*Museum of Natural Science, Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, USA*

[2]*Current address: Department of Biology, Geology and Environmental Science, University of Tennessee at Chattanooga, Chattanooga, Tennessee, USA*

[3]*Museu de Zoologia da Universidade de São Paulo (MZUSP), São Paulo, São Paulo, Brazil*

[4]*Department of Biology, University of Louisiana at Lafayette, Lafayette, Louisiana, USA*

[5]*Department of Ecology and Evolutionary Biology, University of California, Los Angeles, California, USA*

*Corresponding author

Fernando Alda
University of Tennessee at Chattanooga
Dpt. of Biology, Geology and Environmental Science
Holt Hall, 620 E 5th St.
Chattanooga, TN 70803

Phone: (423) 425-4341

E-Mail: alda.fernando@gmail.com

© The Author(s) 2018. Published by Oxford University Press, on behalf of the Society of Systematic Biologists. All rights reserved.
For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

ABSTRACT

Resolving patterns of ancient and rapid diversifications is one of the most challenging tasks in evolutionary biology. These difficulties arise from confusing phylogenetic signals that are associated with the interplay of incomplete lineage sorting and homoplasy. Phylogenomic analyses of hundreds, or even thousands, of loci offer the potential to resolve such contentious relationships. Yet, how much useful phylogenetic information these large data sets contain remains uncertain and often goes untested. Here, we assess the utility of different data filtering approaches to maximize phylogenetic information and minimize noise when reconstructing an ancient radiation of Neotropical electric knifefishes (Order Gymnotiformes) using ultraconserved elements. We found two contrasting hypotheses of gymnotiform evolutionary relationships depending on whether phylogenetic inferences were based on concatenation or coalescent methods. In the first case, all analyses inferred a previously—and commonly—proposed hypothesis, where the family Apteronotidae was found as the sister group to all other gymnotiform families. In contrast, coalescent-based analyses suggested a novel hypothesis where families producing pulse-type (viz., Gymnotidae, Hypopomidae and Rhamphichthyidae) and wave-type electric signals (viz., Apteronotidae, Sternopygidae) were reciprocally monophyletic. Nodal support for this second hypothesis increased when analyzing loci with the highest phylogenetic information content and further increased when data were pruned using targeted filtering methods that maximized phylogenetic informativeness at the deepest nodes of the Gymnotiformes. Bayesian concordance analyses and topology tests of individual gene genealogies demonstrated that the difficulty of resolving this radiation was likely due to high gene-tree incongruences that resulted from incomplete lineage sorting. We show that data filtering reduces gene tree heterogeneity and increases nodal support and consistency of species

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

trees using coalescent methods; however, we failed to observe the same effect when using concatenation methods. Furthermore, the targeted filtering strategies applied here support the use of “gene data interrogation” rather than “gene genealogy interrogation” approaches in phylogenomic analyses, to extract phylogenetic signal from intractable portions of the Tree of Life.

Keywords: coalescence, concatenation, gene tree, Gymnotiformes, ILS, phylogenomics, species tree, UCEs

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

With the advent of phylogenomics and access to large molecular data sets, the systematic community expected unequivocal resolution of the major branches in the Tree of Life (Gee 2003; Rokas et al. 2003). Unfortunately, it is now widely understood that the analysis of hundreds to thousands of genetic loci involves previously unforeseen obstacles. Chief among these is topological discordance among gene trees due to independent evolutionary histories and/or systematic errors in phylogenetic inference (Maddison 1997; Jeffroy et al. 2006; Degnan and Rosenberg 2009; Philippe et al. 2011; Pyron et al. 2014; Doyle et al. 2015; Bravo et al. 2018). As a result, resolving relationships among certain taxonomic groups using genome-wide marker sets has proven challenging, and some fundamental systematic questions remain unresolved or controversial (Chakrabarty et al. 2017; Esselstyn et al. 2017; Reddy et al. 2017).

Some of the hardest problems in phylogenetics involve resolving the sequence of cladogenetic events that occurred during rapidly diversifying radiations. When these events are recent, the short time elapsed since species divergence might be insufficient for genetic drift or selection to sort out alleles before the subsequent accumulation of changes. These rapid divergences may preclude finding a well-supported consensus species tree because of the very shallow genetic differentiation (Pamilo and Nei 1988; Maddison and Knowles 2006). Similarly, rapid and ancient radiations, despite involving species that are highly divergent, can also result in deep coalescence or incomplete lineage sorting (ILS) of alleles (Oliver 2013). Rapid cladogenetic events closer to the root of a tree may cause the stochastic sorting of alleles across the different lineages, which can be further obscured by homoplasy due to saturation. As a result, trees with very short and statistically unsupported branches near the root (i.e., basal polytomies) are frequently observed in rapid and ancient radiations.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Theoretical and simulation studies have shown that concatenation methods of phylogenetic inference, which reduce multi-locus data sets into one single supergene alignment, may produce misleading results when short branches and large effective population sizes coincide (Edwards et al. 2007; Kubatko and Degnan 2007; Degnan and Rosenberg 2009; Roch and Steel 2015). It is under these circumstances that multispecies coalescent methods are preferred, because these methods can accommodate gene tree heterogeneity due to ILS and because these approaches are more likely to recover accurate species trees (Maddison and Knowles 2006; Liu et al. 2015a).

The most commonly used methods to reconstruct species trees under the multispecies coalescent rely on the reconciliation of *a priori* inferred gene trees (i.e., summary methods) (Bravo et al. 2018). Summary coalescent-based methods have, however, been criticized for placing unrealistic confidence in gene tree estimation, and concerns have been raised regarding their performance when the accuracy of inferred gene trees varies (Springer and Gatesy 2016). In some circumstances, such as poor phylogenetic signal or low ILS and high gene tree estimation error, concatenation can provide as good or even better species-tree estimates (Mirarab et al. 2014; Liu et al. 2015a; Rusinko and McPartlon 2017; Molloy and Warnow 2018). Full coalescent methods on the other hand, can overcome these problems by simultaneously estimating gene and species trees and taking into account their uncertainty under a Bayesian framework (Knowles et al. 2012). Unfortunately, these methods are computationally intensive, and they are hard to scale to large data sets (Zimmermann et al. 2014).

It is therefore important to rely on robust gene trees when estimating species trees with summary methods, particularly when attempting to reconstruct ancient relationships or rapid diversifications (Salichos and Rokas 2013). Recently, a number of studies have been published

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

addressing the impact of systematic errors (Meiklejohn et al. 2016; Arcila et al. 2017), the number of loci and their phylogenetic information content (Gilbert et al. 2015; Dornburg et al. 2016; Blom et al. 2017), and different data filtering strategies (Doyle et al. 2015; Hosner et al. 2016; Gilbert et al. 2018; Molloy and Warnow 2018) on gene tree-species tree congruence and phylogenetic reliability.

Filtering genes prior to species tree estimation is commonplace in both concatenation and coalescent-based phylogenomic analyses (Doyle et al. 2015; Hosner et al. 2016; Meiklejohn et al. 2016; Streicher et al. 2016; Blom et al. 2017; Dornburg et al. 2017a; Gilbert et al. 2018; Kuang et al. 2018; Molloy and Warnow 2018). In general, the rationale for these screening criteria is to maximize phylogenetic information while minimizing noise. Loci may be discarded based on the amount of missing data, number of parsimony informative sites, or model fit (Philippe et al. 2005; Doyle et al. 2015; Meiklejohn et al. 2016). These criteria do not, however, consider the structure of the tree (e.g., internode length distribution) or specific regions of the tree when selecting loci, and we refer to these approaches as “non-targeted filtering methods”.

Other methods account for tree structure by estimating phylogenetic informativeness as the probability that a character resolves a particular quartet (Townsend 2007; Dornburg et al. 2017a; 2017b). Phylogenetic informativeness can be calculated along the tree on a per million year basis, and can therefore be used to compare the power of each locus to resolve relationships at different geological time scales. For example, a marker evolving at a certain rate can be of high utility at given tree depths when internodes are long, yet misleading when internodes are short (Su and Townsend 2015; Dornburg et al. 2017b). Based on this principle, such approaches—which we refer to as “targeted filtering methods”—can be used to select loci that

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

are particularly informative for resolving recalcitrant nodes that fall at specific time periods across the Tree of Life (Gilbert et al. 2018).

We carried out a comprehensive study to, first, assess if ILS is a major factor precluding the resolution of the deepest nodes of an ancient rapid radiation of Neotropical freshwater fishes, and second, to evaluate the performance and consistency of different non-targeted and targeted filtering methods when inferring species trees. As a case study, we investigated the relationships among the Neotropical weakly electric fishes of the order Gymnotiformes using ultraconserved elements (UCEs).

Gymnotiformes are thought to have originated during the Lower Cretaceous c. 100-140 Ma on the South American portion of Western Gondwana (Albert 2001; Lavoué et al. 2012; Betancur-R et al. 2017). Currently, there are 254 taxonomically valid extant gymnotiform species in five families: Apteronotidae, Gymnotidae, Hypopomidae, Rhamphichthyidae and Sternopygidae. All species possess electric organs (EOs) that produce weak (mV) electric organ discharges (EODs) used for communication and navigation (Bennett 1971; Albert and Crampton 2005a). EODs are highly stereotyped across species and taxonomic families can be grouped into those species producing signals as discrete electric pulses with intervals of electric silence (e.g., Gymnotidae, Hypopomidae, Rhamphichthyidae) or as a continuous wave (e.g., Apteronotidae, Sternopygidae) (Crampton and Albert 2006). Among these, the electric eel (*Electrophorus electricus*) in the family Gymnotidae is most famous for additionally having a high voltage EO, that can produce discharges of up to 600V (Coates and Cox 1945).

Within the Gymnotiformes, interfamily-level relationships are controversial, and both the taxonomy and systematics of the group have varied during the past decades (see summary in Supplementary Fig. S1). These phylogenetic hypotheses differ mainly in the placement of

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Sternopygidae and Apteronotidae. Morphological studies found varied support for either a sister relationship between these two families (Albert and Campos-da-Paz 1998; Albert 2001), or they resolved Apteronotidae as the sister group to all other Gymnotiformes (Triques 1993; Gayet et al. 1994). Early molecular work based on mitochondrial loci recovered a non-monophyletic Sternopygidae with the genus *Sternopygus* as the sister group to all other gymnotiform families and with the remaining species (there called Eigenmanniidae) forming a clade with Apteronotidae (Alves-Gomes et al. 1995). Additionally, this study considered *Electrophorus* and *Gymnotus* as different families that are sister lineages. Combined approaches using molecular (mitochondrial and nuclear loci) and morphological data recovered a monophyletic Sternopygidae that is sister to Apteronotidae, with Gymnotidae resolved as the sister group to all other species (Albert and Crampton 2005b; Tagliacollo et al. 2016). Notably, these recovered relationships were driven by the inclusion of morphological characters; examining the molecular loci alone failed to resolve the deepest relationships among families and produced a polytomy at the base of Gymnotiformes (Tagliacollo et al. 2016). On the other hand, the most recent and extensive molecular data set proposed a non-monophyletic Gymnotidae with *Electrophorus* recovered as the sister lineage to the remainder of the order, although this study failed to determine the placement of *Gymnotus* with strong support (Janzen 2016).

These studies suggest that traditional Sanger sequencing data sets lack the power to resolve inter-familial relationships within Gymnotiformes, even when taxonomic sampling is relatively complete. Recently, phylogenomic analyses have investigated the relationships of ostariophysan fishes using much larger molecular data sets (e.g., complete mitogenomes, exons, and UCEs). All these studies recovered Hypopomidae and Rhamphichthyidae as the sister clade of Gymnotidae, but differed in finding either Sternopygidae (Elbassiouny et al. 2016; Arcila et

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

al. 2017) or Apteronotidae (Chakrabarty et al. 2017) as their closest relative. Similarly, evolutionary hypotheses inferred using genes that are likely under strong selective pressures, such as globin genes involved in the respiratory system of these fishes (Tian et al. 2017), or genes coding for voltage-gated sodium channels involved in the generation of EODs (Arnegard et al. 2010), recovered either the morphological and total evidence hypothesis of Tagliacollo et al. (2016), with low support, or the molecular hypothesis of Elbassiouny et al. (2016) and Arcila et al. (2017).

Overall, the main conclusion that can be drawn from previous studies is that, except for Rhamphichthyoidea (i.e., Hypopomidae + Rhamphichthyidae), the reconstruction of family-level relationships of Gymnotiformes is highly unstable. These previous studies therefore provide an inconsistent phylogenetic framework in which to evaluate alternative evolutionary hypotheses for the origin and early diversification of the group and its remarkable active electrosensory system. This prior work also illustrates that Neotropical electric fishes provide an excellent opportunity to study different data filtering strategies for inferring species trees in the presence of deep coalescence.

For this purpose, we generated new genomic data for all families and major lineages of gymnotiforms using a custom designed probe set targeting UCEs of Ostariophysi. We inferred species trees based on concatenated and multispecies coalescent methods for data sets differing in the number of taxa and missing data, for which we evaluated the degree of support of the different relationships and the sources of incongruency. We constructed subsets of loci using both non-targeted and targeted filtering methods, and we evaluated the effect of these screening strategies in reducing gene tree heterogeneity and in recovering consistent species trees.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Materials and Methods

Taxon Sampling

Our ingroup is composed of 42 individuals from 39 species across all families of Gymnotiformes. We aimed to have a complete and balanced representation of all major clades and/or phenotypes, and included at least one species from each genus or species group within species-rich genera. The taxon set includes 15 species in 11 genera of Apteronotidae, 11 species in five genera of Sternopygidae, seven species in two genera of Gymnotidae, three species in two genera of Hypopomidae, and three species in three genera of Rhamphichthyidae (Table 1). We included five outgroup taxa representing the three other orders of Otophysi: *Danio rerio* (Cypriniformes), *Astyanax mexicanus* (Characiformes), and *Ictalurus punctatus*, *Hoplosternum littorale* and *Pterygoplichthys sp.* (Siluriformes). Three of these outgroup taxa (*D. rerio*, *A. mexicanus*, and *I. punctatus*) and one *E. electricus* sample are genome-enabled species, therefore their genomes were downloaded from GenBank (GCA_000002035.3, GCA_000372685.1, GCA_001660625.1) and the EFish Genomics database (http://efishgenomics.integrativebiology.msu.edu). All tissue samples in this study belong to vouchered specimens (Table 1), collected and identified by the authors.

Laboratory and Bioinformatic Protocols

For library preparation and targeted enrichment of UCEs we followed the protocols described in Faircloth et al. (2012b, 2013) (Online Appendix 1 in the Dryad Digital Repository https://doi.org/10.5061/dryad.k57430s). We constructed genomic libraries using the Kapa Hyper Prep Kit v.3.15 (Kapa Biosystems) and enriched them using a newly developed set of 6737 probes targeting 2708 UCE loci (Faircloth et al. 2018) following the enrichment protocol for the

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

MYcroarray MYBaits kit v.3.0. We sequenced the libraries in two separate runs of a PE150 Illumina NextSeq 300 (University of Georgia Genomics Facility) and a PE150 Illumina HiSeq 3000 (Oklahoma Medical Research Foundation). Raw read data are archived in the NCBI Sequence Repository Archive (SRA; BioProject ID: PRJNA470726), and concatenated and individual gene alignments are archived in Dryad (https://doi.org/10.5061/dryad.k57430s).

We performed quality control of the demultiplexed sequences, assembled, aligned, and created input files for analyses using programs in the PHYLUCE package (Faircloth 2016) (Online Appendix 2). In addition to the 43 enriched taxa from which we collected new data (Table 1), we extracted the same UCEs *in silico* from the four genome-enabled species (*E. electricus*, *D. rerio*, *A. mexicanus* and *I. punctatus*) using scripts available at https://github.com/carloliveros/uce-scripts (Online Appendix 2).

We created a list containing loci having data for all taxa (a complete data matrix), and another list containing all loci having data for any taxon (an incomplete data matrix). For the complete data matrix, we did not recover any UCE locus shared across all taxa. Therefore, we constructed a reduced taxon set of eight ingroup taxa representing all gymnotiform families and three outgroups (Table 1). We selected two species from each major clade in each family that also maximized the number of shared loci recovered, except for Hypopomidae and Rhamphichthyidae, for which we captured the largest number of UCEs and we included one representative species per family. This reduced taxon set resulted in a complete data set of 368 loci. For the incomplete matrix of the full taxonomic set, we filtered all loci to create a matrix that was at least 75% complete (alignments contained ≥35 of 47 individuals) using *phyluce_align_get_only_loci_with_min_taxa.py*, resulting in 966 UCE loci. For comparative purposes, and to assess the relative effect of number of taxa and missing data, we also created a

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

75% complete matrix for the reduced taxon set, which contained 1472 UCE loci (Online Appendix 2).

Analysis of Concatenated UCE Data

We estimated best-fit partitioning schemes for all our concatenated matrices using the Sliding-Window Site Characteristics (SWSC) method described in Tagliacollo and Lanfear (2018). SWSC uses a sliding-window approach to divide each UCE into data blocks with similar characteristics, such as site entropy, to generate partitions that account for heterogeneity in rates and patterns of molecular evolution within each UCE locus (code available from https://github.com/Tagliacollo/PFinderUCE-SWSC-EN). We then used this output to estimate the optimal partition scheme of the whole alignment by grouping together similar data blocks using PartitionFinder 2 (Lanfear et al. 2016). Due to computational limitations, we only considered the GTR + G model of molecular evolution for each subset, and we compared possible schemes using the relaxed hierarchical clustering algorithm variant (rclusterf) (Lanfear et al. 2014) and the AICc criterion. The best scheme for the 100% complete reduced taxon set contained 259 partitions (1-20 subsets, 50-2838 bp per partition), and the 75% complete full and reduced taxon sets had 544 (1-20 subsets, and 50-3338 bp per partition) and 660 partitions, respectively (1-26 subsets, and 50-3777 bp per partition).

We conducted 20 Maximum Likelihood (ML) searches for the phylogenetic tree that best fit the data under the GTRGAMMA substitution model in RAxML v.8.0.19 (Stamatakis 2014) (Fig. 1). We used non-parametric bootstrapping to assess nodal support by allowing the program to automatically determine the number of replicates using the autoMRE criterion. Bootstrap

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

searches stopped after 60 replicates in all analyses. We ran the analyses using a non-partitioned matrix and the best-fitting partitioning schemes described above.

We also estimated phylogenetic trees for our concatenated data sets in a Bayesian inference (BI) framework using the MPI version of ExaBayes v.1.4.1 (Aberer et al. 2014) (Fig. 1). For each analysis, we ran four independent MCMCMC runs for 3 x 10^6 generations, composed of four coupled chains (one cold and three heated chains), and a sampling frequency of one tree every 500 generations. We assessed convergence based on visualizations of the scale reduction factors and the average deviation of splits frequencies using the postProcParam and sdsf programs in ExaBayes v.1.4.1, and the effective sampling sizes (ESS) values obtained for all estimated parameters in Tracer v.1.6.0 (Rambaut et al. 2014). We discarded 25% trees as burn-in.

Species Tree Analysis

To account for gene tree incongruences due to ILS, we inferred species trees using the summary coalescent-based method of ASTRAL-II (Mirarab and Warnow 2015) (Fig. 1). ASTRAL-II uses a quartet-based approach to estimate species trees from a set of gene trees, so first we estimated gene trees for each UCE locus by conducting 20 ML searches in RAxML v.8.0.19 under the GTRGAMMA substitution model, and assessed support by independently generating 200 non-parametric bootstrap replicates of each alignment. Then, we used the best fitting ML gene trees as input and conducted gene + site resampling on the bootstrap replicates.

We also co-estimated gene trees and the species tree in a Bayesian framework using a full coalescent method in *BEAST (Heled and Drummond 2008). Because *BEAST is prohibitively computationally intensive for such a large data set, it was not feasible to analyze all loci and

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

taxa. Hence, we conducted four replicated analyses using subsets of 50 loci randomly sampled from the reduced taxon set with no missing data (i.e., eight ingroup and three outgroup taxa, using 50 loci from a total of 368 loci) (Fig. 1). We used the HKY substitution model and empirical base frequencies for all loci, and a Yule species tree prior with a linear-with-constant-root population size model. We ran the analyses for 2 x 10^8 generations with sampling every 2 x 10^4 generations. We analyzed the output species trees from *BEAST using Densitree2 (Bouckaert and Heled 2014).

We conducted Shimodaira-Hasegawa (SH) tests to examine statistical differences between tree topologies obtained using concatenation and coalescent-based methods for both full and reduced taxon sets, with and without missing data, in RAxML v.8.0.19.

Gene Tree-Species Tree Discordances

We calculated the amount of gene tree discordance using frequentist and Bayesian methods. First, we calculated the frequency of ML gene trees that recovered each family as monophyletic, using the program ape v.4.1 (Paradis et al. 2004) in R 3.3.3 (R Core Team 2017). We also used the R package phangorn v.2.2.0 (Schliep 2011) to evaluate topological discordances between all trees using Robinson-Foulds (RF) distances (Fig. 1), which calculates the number of internal splits that differ between two trees (Robinson and Foulds 1981). To compare trees with different numbers of taxa, we calculated the relative differences in RF distances by calculating the ratio of the RF distance divided by the maximum RF distance (%RF), which is defined as $2(n–2)$, where n is the number of taxa, and $n–2$ is the maximum number of inner branches for a rooted tree (Kuhner and Yamato 2015). To further evaluate how well the multispecies coalescent can explain the observed gene tree variation we used the species

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

trees inferred in ASTRAL-II for each of our data sets, with branch lengths in coalescent units, to simulate 10,000 gene trees under the coalescent model using the R program phybase v.1.5 (Liu and Yu 2010). We sampled 966, 1472 and 368 gene trees, corresponding to the number of loci in the complete data set and in the reduced data sets with and without missing data, that served as the gene trees expected from the coalescent model for each of our data sets. We then calculated RF distances and their respective %RF for the expected gene trees to measure the expected gene tree variation, and compared these to the observed gene tree variation calculated above.

Second, we carried out a Bayesian concordance analysis in BUCKy v.1.4.4 (Ané et al. 2007; Larget et al. 2010) to calculate concordance factors (CF), representing the proportion of gene trees in which given bipartitions are present. BUCKy assembles taxa with highest CF values into clades with the least conflict in order to generate a primary concordance tree. This program also applies CFs in a quartet-joining algorithm to construct a population tree, which is consistent with a coalescent framework species tree when gene tree discordance is due to ILS. We used the posterior distribution of trees obtained from the ExaBayes analysis, ran BUCKy for 1 x 10^6 generations on four chains, and discarded 1 x 10^5 generations as burn-in. We repeated the analyses with increasing levels of *a priori* discordance among gene trees setting the alpha prior to α=1, 5, and 10. We also used the independent prior, which is equivalent to $\alpha=\infty$, and assumes that all gene tree topologies are independent from each other. Both the number of taxa and the amount of missing data can affect the estimated CFs, with more taxa and more missing data generally leading to lower CFs (Ané 2013). Based on this knowledge, we aimed to minimize the number of individuals and maximize the number of loci that informed CF estimation by only using the reduced taxon set with no missing data for this analysis (Fig. 1).

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

To explore the causes of gene tree incongruences we employed a "gene genealogy interrogation" (GGI) approach (Arcila et al. 2017) that is designed to discern if genealogical discordance is due to ILS or to systematic error associated, for example, with low phylogenetic signal in the data or model misspecifications. Briefly, GGI carries out approximately unbiased (AU) topology tests for each gene alignment to evaluate their relative support of alternative species tree hypotheses based on partially constrained trees. The best-fit hypotheses can then be ranked for each gene, and the highest ranked gene trees can be used in a summary species-tree analysis (e.g., ASTRAL-II). In our case, we were interested in resolving the deepest relationships among families, which were the source of the incongruence. Therefore, we defined 105 hypotheses representing all possible rooted trees for five lineages (i.e., all gymnotiform families), and one additional "null hypothesis" where relationships among lineages were not resolved and therefore represented as a polytomy at the crown of Gymnotiformes. We then conducted 102,396 constrained ML searches for the full taxon set (Fig. 1), where each family was constrained to be monophyletic, except Gymnotidae, in which *E. electricus* was not forced to be the sister taxon to *Gymnotus*, as this relationship seemed unstable across our analyses (see Results) and previous studies (Janzen 2016). Also, to retain a feasible number of hypotheses to test (e.g., 105 possible rooted trees for five lineages vs. 945 trees for six lineages), we kept Hypopomidae + Rhamphichthyidae (i.e., Rhamphichthyoidea) as a clade because they were unambiguously recovered as sister groups in all previous—as well as our own—phylogenetic reconstructions (Triques 1993; Alves-Gomes et al. 1995; Albert and Campos-da-Paz 1998; Tagliacollo et al. 2016). No other constrains on the relationships within families or on branch lengths were imposed for the GGI approach.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Despite there being strong support for the monophyly of the families and lineages that we constrained (e.g., morphological, mitochondrial, multi-locus), it is possible that some gene trees show instances of deep coalescence and do not recover the monophyly of some subclades. The GGI procedure, however, assumes that all lineages are monophyletic and as a result non-sorted allelic polymorphisms can bias the results of the analyses (Arcila et al. 2017). Therefore, we tested whether the assumption of monophyly is met by the major gymnotiform groups based on theory predicting that under a neutral coalescent model >99.99% of genes in a genome will achieve monophyly after 5.3−8.3 coalescence time units (Rosenberg 2003). We used a fossil-calibrated phylogeny (see below) and converted branch lengths in millions of years (*T*) into coalescent units (τ) using the formula $\tau = T/(2N_e \bullet \text{generation time})$. Using a range of effective population sizes (N_e=10,000 and 100,000) and generation times (2.5 and 5 years) (Mims et al. 2010; Arcila et al. 2017), the estimated branch lengths for the gymnotiform groups of interest were between 16.22−38.42 coalescent units. These coalescent times are consistent with estimates for the stem length of all Gymnotiformes being between 20 and 122 coalescent units (Arcila et al. 2017), and suggest that sufficient time has elapsed for gene trees to achieve subclade monophyly.

Data Filtering

A common approach to reducing gene tree heterogeneity and to increase global support is to follow a “gene data interrogation” approach and select loci that are expected to provide the highest phylogenetic signal (Philippe et al. 2005; Doyle et al. 2015; Gilbert et al. 2018). A straightforward method for filtering subsets of data is by using the number of parsimony informative characters as a proxy for phylogenetic information content (Hosner et al. 2016;

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Meiklejohn et al. 2016). For each taxon set (reduced and full) and level of matrix coverage (100% and 75% complete), we calculated the number of parsimony informative sites for each locus using *phyluce_align_get_informative_sites_per_taxon.py.* Another filtering approach that has proven useful in improving phylogenetic congruence and reliability is to remove rapidly evolving genes (e.g., genes deviating from the molecular clock; Doyle et al. 2015), because these are more likely to result in biased inferences under simple substitution models, or in other artifacts such as long branch attraction (Felsenstein 1978; Brinkmann et al. 2005; Nozaki et al. 2007; Zhong et al. 2011). We used the likelihood ratios (LRTs) from molecular clock tests as a relative measure of clock-likeness across genes. Given the selected model of sequence evolution and the ML tree for each gene, we calculated likelihoods twice: once enforcing a strict clock model and once estimating each branch length independently. In both cases, we calculated likelihoods with the programs ape v.4.1 and phangorn v.2.2.0 in R, using parameter and branch length estimates from RAxML. We calculated LRTs between these models as twice the difference in their log-likelihood scores and then sorted the genes in ascending order by their LRT.

In addition to these two common strategies of non-targeted phylogenomic filtering, we used two alternative targeted filtering approaches that are based on estimates of phylogenetic informativeness. Phylogenetic Informativeness (PI) calculates the probability that a character resolves a hypothetical polytomy in a four-taxon phylogeny and then remains unchanged along the peripheral branches (Townsend 2007). The shape of PI profiles can be used to predict the utility of sequences for inferring relationships across entire topologies as well as for individual nodes (Townsend 2007; Dornburg et al. 2016). Therefore, in contrast to non-targeted filtering, PI-based methods of targeted filtering do not use point estimate statistics as proxies to determine

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

phylogenetic information content *a posteriori*. Rather PI-based approaches take into account the phylogeny and the timescale of the relationship(s) of interest.

Prior to inferring PI, we constructed a time-calibrated phylogeny of lineages upon which PI calculations are dependent. To calibrate the chronogram, we used the oldest fossil of crown-group Ostariophysi, which is the stem gonorynchiform †*Rubiesichthys gregalis* (140-145.5 Ma) (Fara et al. 2010), and a secondary calibration derived from geological and fossil data (e.g., the Miocene fossil †*Humboldtichthys kirschbaumi*, Gayet et al. 1994) for the crown of Gymnotiformes (67-118 Ma) (Tagliacollo 2015). We specified hard lower bounds, representing the respective calibration ages, and we inferred a timetree using the penalized likelihood method and the non-correlated rates of molecular substitution ("relaxed") model (Paradis 2013) implemented in ape v.4.1. According to this analysis, our estimated time for the deepest divergence among families of Gymnotiformes was congruent with previous studies, placing it between 80−100 Ma (Albert 2001; Lavoué et al. 2012; Betancur-R et al. 2017).

We used the program TAPIR (Pond et al. 2005; Townsend 2007; Faircloth et al. 2012a) to calculate PI of each UCE locus. TAPIR calculates substitution rates based on the best estimated substitution model for each locus and then uses those rates to estimate the PI profile of each locus (Faircloth et al. 2012a). We calculated PI per nucleotide per locus per data set, we used the program PhyDesign (López-Giráldez and Townsend 2011) to integrate these values (i.e., calculate the area under the curve) for the epoch of interest (i.e., the crown node of Gymnotiformes, 80−100 Ma), and we ranked loci based on the PI for that epoch. Integrated values will be larger for those genes that have a higher probability of exhibiting substitutions in the given epoch that will not be obscured in subsequent branches.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

However, integration of PI does not account for phylogenetic noise or how homoplastic site patterns from fast-evolving sites may influence phylogenetic resolution of specific nodes. Phylogenetic Informativeness should perform well for slow evolving genes and/or shallow time scales; however, even small amounts of homoplasy can overwhelm the phylogenetic signal when attempting to resolve short and deep internodes (Townsend et al. 2012).To account for this effect, we calculated phylogenetic signal relative to phylogenetic noise using quartet internode resolution probabilities (QIRP) (Townsend 2007; Townsend and Lopez-Giraldez 2010; Townsend et al. 2012), which are based on a predictive relationship between the probability of resolving a given node and the rate of evolution, internode distances, and tree depth (Townsend et al. 2012). Similarly to PI, we used PhyDesign to integrate nucleotide QIRP values and to estimate the probability of each locus to resolve a polytomy at the crown of Gymnotiformes.

Following each of these four filtering criteria, we ranked all loci and binned them into quartiles to create data subsets using the first upper quartile (25%), and the two upper quartiles (50%), such that the reduced taxon set with no missing data contained 97 UCEs for the top 25% and 184 loci for the top 50%, the full taxon set contained 241 and 483 loci for the top 25% and 50% loci, and the reduced taxon set with 25% missing data contained 368 and 736 loci. We then carried out concatenated analyses (i.e., ML in RAxML and BI in ExaBayes), and inferred multi-species coalescent species trees in ASTRAL-II for the 24 filtered data subsets of most informative loci as described above (Fig. 1).

Finally, for each of the different filtering strategies, we assessed their performance in reducing gene tree discordance by recalculating and comparing the proportion of gene trees for which each family was monophyletic, RF distances, and CFs from BUCKy (Fig. 1). We also tested the effect of data filtering on locus length and on reducing base compositional bias. We

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

carried out *t*-tests to test differences in length between loci in the non-filtered and filtered data sets, and tested the correlation between each filtering criterion and locus length using Pearson's correlation in R. Also, we performed Chi-squared tests of homogeneity of state frequencies across taxa and families for all filtered and non-filtered data sets in PAUP* v4.0a163 (Swofford 2003). We calculated the GC content for each species across all data sets as a percentage of the total alignment length, and tested for significant differences in GC composition among filtered and non-filtered data sets using a Linear-Mixed Effect model with individuals as random factors in the R package nlme 3.1 (Pinheiro et al. 2018).

RESULTS

Phylogenomics of Gymnotiformes Using Concatenated and Coalescent Methods

We collected a total of 2681 UCE loci from 44 samples using a hybrid capture and enrichment protocol and from three published genomes *in silico*. We sequenced an average of 1,603,275.71 reads per sample (min-max=356,056-4,418,216), which we assembled into 21,246.69 contigs (min-max=1736-149,135) per sample. On average, we recovered 1499.02 UCE loci (min-max=985-2205) per sample that produced alignments 473.83 bp long (min-max=111-2518). Each locus was sequenced for 27.30 individuals (56.9%) on average, but some loci were recovered for as few as three individuals. Therefore, in our alignments, we only used those loci that included at least 75% of the individuals (i.e., a minimum of 35 individuals per locus). This added up to 966 UCE loci and a total of 376,533 bp. We also constructed a reduced taxon set with no missing data for eight ingroup and three outgroup taxa that included 368 UCE loci and 170,794 bp, and the same reduced taxon set allowing for 25% missing data, which included 1472 UCEs and 553,207 bp (Fig. 1).

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Our concatenated analyses produced consistent results across all data sets and methods (e.g., ML and BI, with data partitioned and non-partitioned). As a conservative measure of support, we only report results from the partitioned analyses, since these provided slightly lower bootstrap values. These analyses recovered all families as monophyletic with high support (Bootstrap support: BS=100; Posterior probability: PP=1.00), except for Gymnotidae, which was not recovered as monophyletic in the reduced taxon set with no missing data. However, the analysis of the reduced taxon set with 25% missing data, and of the full taxon set, recovered *E. electricus* as the sister group to *Gymnotus* (BS=95, PP=1.00, and BS=73, PP=1.00, respectively) (Fig. 2, Supplementary Fig. S2). The relationships among families were the same for the full and reduced taxon sets with 25% and no missing data. In all cases, Apteronotidae was the sister group to all the other families (BS≥78, PP=1.00), and Sternopygidae was the sister group to a clade that included Gymnotidae, Rhamphichthyidae and Hypopomidae, the later two of which were each other's sister group (Fig. 2, Supplementary Fig. S2). In the analysis of the full taxon set, relationships within families were also highly supported (mean BS=99.22±4.07; mean PP=1.00±0.00).

The species trees inferred using ASTRAL-II recovered the same relationships within families as in the concatenated analyses with high support. The only exception was in the Gymnotidae, where neither the analysis of the full taxon set, nor the reduced taxon sets, recovered with high confidence the phylogenetic position of *E. electricus* (BS=55-74; Fig. 3, Supplementary Fig. S3) within the clade including *Gymnotus*, Hypopomidae and Rhamphichthyidae. Furthermore, this later clade was not recovered as the sister group to Sternopygidae, as it was in the concatenated analyses, but rather as the sister group to a clade

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

formed by Sternopygidae and Apteronotidae (support for this relationship ranged between BS=58-71 depending on the data set analyzed; Fig. 3, Supplementary Fig. S3).

The SH test for the full taxon set showed that the species tree topology was significantly less likely than the concatenated tree hypothesis (ln L=-2,710,018.27 vs. ln L=-2,709,971.88, P<0.05, respectively). For the reduced taxon sets, both with (ln L=-2,459,894.24 vs. ln L=-2,459,919.42) and without missing data (ln L=-847,442.39 vs. ln L=-847,458.65), the SH tests did not show a significant difference between the likelihood scores of the best ML trees recovered and the ML phylogeny constraining Apteronotidae + Sternopygidae. This result may suggest a larger influence of the number of taxa than the number of loci in the relationships recovered.

In the full coalescent method implemented in *BEAST, the replicated analyses of the reduced taxon data set produced between one and nine consensus species trees. In all subsets, the most common tree (87.08%-100% of the trees) recovered two reciprocally monophyletic groups formed by Sternopygidae + Apteronotidae and by Gymnotidae + (Rhamphichthyidae + Hypopomidae), respectively (Fig. 4). Also, the most common trees always recovered a sister relationship between *E. electricus* and *Gymnotus*, and only 0.14% of the trees did not recover this relationship. Similarly, in only 7.54% of the trees was Apteronotidae recovered as the sister group to all the other families. The primary concordance and population trees obtained in the BUCKy analyses that assumed low to moderate levels of gene tree discordance (α=1, 5, 10, equivalent to 6 to 50 gene tree clusters) agreed in recovering Apteronotidae as the sister group to other gymnotiform families, and Sternopygidae as the sister group to the clade formed by Gymnotidae + (Rhamphichthyidae + Hypopomidae). However, in the analyses that did not consider any gene clusters (i.e., independent gene trees, $\alpha=\infty$), the population trees differed from

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

the primary concordance trees in recovering Apteronotidae as the sister group to Sternopygidae, with the other relationships among and within families remaining the same (Supplementary Fig. S4).

Gene Tree and Species Tree Disagreement

Individual gene trees showed large disagreement in their topologies, and in the degree of support for different relationships. For example, whereas 92.75% of the gene trees recovered the monophyly of Apteronotidae, only 21.33% recovered Gymnotidae as monophyletic (i.e., *E. electricus* as the sister taxon to *Gymnotus* species), and only 14.18% recovered Apteronotidae sister to Sternopygidae (Table 2). These differences were also made clear by the large %RF distances among all gene trees, which ranged between 86.10% for the full taxon set and 55.78% for the reduced taxon set with no missing data. Furthermore, it is remarkable that in all data sets the heterogeneity observed among gene trees was significantly higher than the variation expected under a neutral coalescent model, as estimated through simulations based on the inferred species trees (Supplementary Fig. S5). This result is elaborated on by the fact that the coalescent model accounts for 41.43−52.20% of gene tree variation observed in the real data sets of 966 and 368 loci, but for only 9.99% of the gene tree variation in the 75% complete reduced taxon set with 1472 loci.

In the Bayesian concordance analysis performed in BUCKy, CFs were highest for the monophyly of the Gymnotiformes (CF=0.96, 95%CI=0.95-0.97) and for some terminal groups (e.g., Apteronotidae CF=0.98, 95%CI=0.97-0.99, Rhamphichthyoidea CF=0.91, 95%CI=0.88-0.92), but much lower for the relationships among families as well as for the monophyly of Gymnotidae (CF=0.28, 95%CI=0.24-0.33) (Supplementary Table S1). In general, CFs decreased

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

with higher levels of *a priori* assumed gene tree discordance and, consequently, were lowest when α=∞. Interestingly, for the highest α value, the CF credibility intervals for the conflicting splits between the primary concordance and the population trees overlapped with all their alternative topologies, a pattern consistent with ILS as the only cause of discordance. However, we did not observe this pattern in the analyses with lower values of α (Supplementary Table S1).

To assess if the observed differences among gene trees were due to ILS or to systematic error, we carried out a GGI analysis. Among the 105 rooted tree hypotheses that we tested, we found that the most frequent best fit tree corresponded to the phylogenetic hypothesis recovered in all our concatenated analyses, but this was only found in 6.93% of our loci (67 out of 966 loci). Out of these 67 trees, only two were statistically better than the second ranked tree. The hypothesis proposed by our species tree analyses was recovered as the best fit tree in 28 of the analyzed loci (2.89%), of which four were significantly better than the alternative topology with the second highest score. Remarkably, 104 out of the 105 tested hypotheses were recovered as the best fit tree for at least one of our UCE loci, but only 30 of them were the best fit for more than 1% of the gene trees (Supplementary Fig. S6a). When we carried out the GGI analysis including the null hypothesis where all families were placed as a polytomy at the crown of the Gymnotiformes, this topology showed the best fit score for the majority of loci (466 out 966 loci, 48%), suggesting that individual genes contain insufficient phylogenetic information to support any of the proposed hypotheses (Supplementary Fig. S6b).

Effect of Data Filtering

The concatenated analyses of all the filtered data sets produced the same topology as the non-filtered data, and in all cases the sister relationship of Apteronotidae and all the other

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

families of Gymnotiformes was highly supported. However, we observed a number of differences in the species trees inferred from filtered data using ASTRAL-II. In general, the analysis of the filtered data from the full taxon set increased the support for the placement of Sternopygidae sister to Apteronotidae and *E. electricus* sister to *Gymnotus* (See Table S2 for a summary of bootstrap values of the major nodes in the species trees inferred for all data sets, and Supplementary Fig. S7 for a summary of all phylogenetic hypotheses inferred across all methods and data sets). The only exception was the filtered data set including the first and the two upper quartiles of the most clock-like UCE loci, which rendered lower support than the non-filtered data for the sister relationship of Apteronotidae and Sternopygidae (Fig. 3, Supplementary Table S2, Figs. S8-S9). On the other hand, filtering the data of the reduced taxon sets only increased the support for the Apteronotidae + Sternopygidae relationship in five out of the eight species trees inferred using alignments with up to 25% missing data, and in only one of the species trees constructed with no missing data (Supplementary Figs. S10-S13). Also, the species trees inferred using the first and the two upper quartiles of loci with the most parsimony informative sites did not recover *E. electricus* sister to *Gymnotus*, and the first upper quartile of loci with highest phylogenetic informativeness placed Apteronotidae sister to Sternopygidae (Supplementary Table S2, Figs. S10a and S13c).

All species trees inferred in *BEAST using filtered data from the reduced taxon set agreed on their most common topology, which was the same as for the non-filtered data (Supplementary Fig. S7). Uncertainty—represented by the number of consensus trees—was lowest when data sets were pruned based on phylogenetic informativeness and QIRP, both of which rendered a maximum of three consensus trees. The most common topology when filtering according to QIRP was recovered in 97.63% of the trees and in 88.7% of the trees filtered by

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

phylogenetic informativeness. When loci were screened based on clock-likeness, uncertainty for some of the replicates was higher than in the non-filtered data. In those replicated analyses, there were a number of low frequency species trees (<10%) that recovered the same relationship that was highly supported in the concatenated analyses.

Following data filtering, we observed an increase in the percentage of trees recovering the monophyly of all families in all data subsets except for the most clock-like loci. We observed the largest increase in the data sets filtered according to their QIRP, which increased by >10% the agreement of gene trees recovering Rhamphichthyidae (81.26%) and Sternopygidae (52.38%) as monophyletic. However, this same strategy only increased by 3.73% the number of gene trees with monophyletic Gymnotidae (25.05%), despite this was the largest improvement among all (Table 2).

Similarly, the filtering strategies for which we observed the largest reduction (Δ) in %RF were the first upper quartile of loci with highest QIRP, phylogenetic informativeness and parsimony informative sites in the full taxon set (Δ%RF=0.05-0.6%) and in the reduced taxon set allowing for 25% of missing data (Δ%RF=0.11-0.12%). Interestingly, the molecular clock criterion performed worst among all filtering methods—both when considering the first and the two upper quartiles (Fig. 5, Supplementary Table S2).

Concordance factors were always higher for the filtered than for the non-filtered data sets. In the analyses with $\alpha=\infty$, the highest increase in CF across all splits was observed for the QIRP criterion, with few exceptions, such as the split of Sternopygidae and the group including Gymnotidae, Hypopomidae and Rhamphichthyidae, that was slightly higher when we selected loci based on their parsimony informative sites (CF=0.18 and 0.19, respectively). Filtering, however, did not change the primary concordance and population trees (Supplementary Fig. S4),

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

except when analyzing the most clock-like loci. Here, the primary concordance tree resolved a non-monophyletic Gymnotidae, where *E. electricus* was the closest relative of the Rhamphichthyoidea, although the CF for this split (CF=0.19, 95%CI=0.15-0.23) was only slightly higher than for the most commonly recovered monophyletic Gymnotidae (CF=0.19, 95%CI=0.14-0.23). As in the analyses of the non-filtered data, population trees differed from the primary concordance trees in recovering Sternopygidae sister to Apteronotidae. The highest CF for this relationship was obtained after analyzing loci with the highest number of parsimony informative sites and highest QIRP (CF=0.15, 95%CI=0.11-0.20).

Pruning of data resulted in a significant increase in the length of the aligned loci (Student's *t*-test *t*=25.19–32.70, *P*<0.001). Also, all the data sets showed a significant and positive correlation between their measures of phylogenetic information content and locus length (Pearson's *r*=0.74–0.88, *P*<0.001), except for the most clock-like loci that showed a significant, although weak, negative correlation (Pearson's *r*=-0.11, *P*=0.001) (Supplementary Fig. S14). Alignments of the most clock-like loci were also significantly shorter than loci that were filtered out (mean length=342.66 bp compared to 437.42 bp, Student's *t*-test *t*=-12.27, *P*<0.001). These results are intuitive because shorter loci may represent those with lower evolutionary rates and are expected to contain fewer informative sites, which ultimately should result in higher estimation errors of individual gene trees that, if discordant, will be eliminated by our filtering methods (Simmons 2014; Hosner et al. 2016).

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Base composition significantly deviated from homogeneity across all species and families in the non-filtered full taxon set, with the exception of Apteronotidae and Hypopomidae (Supplementary Table S3). Data filtering decreased the Chi-square values, indicating a better fit of the observed to the expected data, but all tests remained significant for all comparisons except in the Gymnotidae (Supplementary Table S3). All filtered data sets had significantly lower GC content than the non-filtered data sets, and for each filtering criterion the first quartile loci had significantly lower GC content than the data sets including first and second quartiles (Supplementary Table S4). Once again, the exception was the data set containing the first quartile of most clock-like loci, which did not differ significantly in GC content from the non-filtered data set, and had a significantly higher GC content than the first and second quartile locus data set. Finally, the first and second quartiles of loci with the highest QIRP values always showed the lowest overall percent of GC across all taxa and families (Supplementary Table S4, Fig. S15).

DISCUSSION

In this study, we investigated the basis of gene tree-species tree heterogeneity and explored strategies to reduce these incongruences and recover accurate species trees in an ancient and rapid radiation of Neotropical freshwater fishes. We used genomic data from hybrid capture enrichment of UCEs and carried out extensive phylogenomic analyses using a combination of concatenated and coalescent-based methods. We recovered two recurring topologies, one with high support in all concatenated analyses, and another topology with moderate support in the coalescent-based analyses. Heterogeneity of gene trees was extremely high and showed patterns that were compatible with low phylogenetic informativeness of individual loci, and with high

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

ILS being the main source for these incongruences. Data filtering, and particularly targeted filtering approaches, were effective in increasing consistency among gene trees and support for the species tree hypothesis without the need to constrain or select hypotheses *a priori*. On the other hand, concatenated analyses were insensitive to the use of any of the data filtering strategies tested.

Based on our most rigorous and complete findings, we propose a novel species tree hypothesis for the Gymnotiformes where the families generating pulse-type (Pulseoidea) and wave-type (Sinusoidea) electric signals are reciprocally monophyletic. Associated implications for the evolutionary history of the electric organs in Neotropical weakly electric fishes are discussed below.

Phylogenomics of Gymnotiformes

We inferred a phylogenomic hypothesis that recovered the monophyly of the five families of Gymnotiformes with high support and consistently across methods. Relationships among genera within families were also highly congruent for all types of analyses—concatenated and coalescent-based, non-filtered and filtered—and with respect to the most recent molecular phylogenetic hypotheses (Janzen 2016; Tagliacollo et al. 2016).

Of notably exception are the relationships within Gymnotidae. While our results were generally congruent with previous works, both in the relationships among species of *Gymnotus*, and in recovering the monophyly of the family, Gymnotidae was the family least frequently recovered as monophyletic by each individual gene tree (between 21.33%-25.73%, before and after filtering the data). Similarly, only two out of nine (22.2%) genes analyzed in Janzen (2016) recovered *E. electricus* as the sister group to *Gymnotus*, and as in our analyses, *E. electricus*

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

generally showed very long terminal and short internal branches. When this occurs in the true species tree, long branches may cause artefactual phylogenetic groupings due to an inherent bias in the estimation procedure (Hendy and Penny 1989). This bias can further be exacerbated when rates of evolution vary (Maley and Marshall 1998), as could be the case under the strong selective pressures to which this unusual species may have been exposed (Tian et al. 2017). The electric eel (*E. electricus*) is characterized by a large number of autoapomorphies, including unique electrogenerative and respiratory traits (Johansen et al. 1968; Moller 1995; Albert and Campos-da-Paz 1998). In any case, relationships involving *E. electricus* were generally recovered with only moderate or low support in both concatenated and coalescent analyses. This species has been previously considered as a monotypic family or as the sister group to all other gymnotiforms (Triques 1993; Gayet et al. 1994; Janzen 2016); that relationship was never recovered in our analyses. Instead, the second most frequently recovered bipartition, but with CF credible intervals overlapping with the *E. electricus* + *Gymnotus* split, was *E. electricus* as the sister clade to Rhamphichthyoidea. That relationship has never been proposed in previous morphological or molecular studies.

In contrast to the highly congruent intra-familial relationships, among-family relationships differed strikingly across different phylogenetic methods and from previously proposed hypotheses. In brief, concatenation consistently recovered Apteronotidae as the sister group to all other Gymnotiformes with high support, which is also a topology that has commonly been recovered by molecular and morphological analyses (Triques 1993; Gayet et al. 1994; Arnegard et al. 2010; Elbassiouny et al. 2016; Arcila et al. 2017). Based on morphology, this relationship supports the presence of a caudal fin in Apteronotidae being plesiomorphic, since only members of this group and *E. electricus* possess it–although the occurrences of

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

these morphologically equivalent tail structures have been proposed to have evolved independently (de Santana et al. 2013). Molecular phylogenies that have recovered this relationship (Apteronotidae as the sister group to all other Gymnotiformes) are either problematically incomplete (in terms of their sampling of Gymnotiformes) because they are part of broader scale phylogenetic analyses (Arcila et al. 2017; Betancur-R et al. 2017), or as in our case, use concatenated phylogenetic methods (Elbassiouny et al. 2016). The main conclusion that can be drawn from this hypothesis is that it proposes a monophyletic Pulseoidea (i.e., families generating pulse-type EODs: Gymnotidae, Hypopomidae and Rhamphichthyidae), but not Sinusoidea (i.e., families generating wave-type EODs: Apteronotidae and Sternopygidae), which is paraphyletic with respect to pulse-type EOD families, and therefore the latter could be considered as the derived state (see below: *Evolutionary History of Electric Organs and Signals*).

On the other hand, coalescent-based methods rendered species trees where Pulseoidea and Sinusoidea were reciprocally monophyletic. Although previous morphological and molecular phylogenetic studies had proposed the monophyly of either one of these groups, none had formally hypothesized an evolutionary scenario where both groups were monophyletic. So far, only Janzen (2016) has proposed reciprocally monophyletic Apteronotidae + Sternopygidae and Rhamphichthyoidea + *Gymnotus*, but that study did not recover the monophyly of Gymnotidae. The bootstrap support for our novel hypothesis was low in the ASTRAL-II analysis using the full taxon set, but it was consistently resolved across all data sets. Highest support was obtained when including the largest number of loci in the reduced taxon sets and allowing for 25% missing data (Supplementary Table S2). This result would be in agreement with empirical studies suggesting that the number of loci, despite potentially adding noise into the analysis, is a

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

more important factor for recovering stable relationships using coalescent based methods than, for example, the amount of missing data (Hosner et al. 2016; Streicher et al. 2016; Molloy and Warnow 2018). Most significantly, the full coalescent model implemented in *BEAST, which presumably provides the most accurate species tree and should be preferred over summary coalescent methods (Knowles et al. 2012), always recovered the Pulseoidea + Sinusoidea hypothesis with high support.

Difficulties Inferring an Ancient and Rapid Radiation

Our results bear out the difficulty of other studies in recovering a robust and consistent phylogenetic hypothesis for the Gymnotiformes. These difficulties are mainly due to the very short deep branches suggesting short speciation intervals that, on one hand, provide little time for substitutions to accumulate on internal branches (i.e., low informativeness) and, on the other hand, increase the probability of ILS to occur and consequently drive the observed incongruences (McCormack et al. 2013; Suh et al. 2015).

Further evidence for this assumption was provided by the Bayesian concordance analyses in BUCKy, where in most cases we observed overlapping frequencies of conflicting splits among families (Supplementary Table S1). When gene tree heterogeneity is mostly caused by ILS, the multispecies coalescent model predicts gene trees that are equally frequent for all triplets of species, hence the equal CFs observed (Degnan and Rosenberg 2009; Chung and Ané 2011). In these situations, the topology obtained using coalescent methods is presumably more accurate than from concatenation—which essentially can be considered as a particular case of the multispecies coalescent where all gene trees are evolving under the same topology (an unlikely proposition). However, it is also important to note that coalescent methods assume that gene tree

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

heterogeneity is only due to ILS, rather than from analytical errors. According to our simulations, the neutral coalescent model may account for about 50% of the observed gene tree variation, therefore there can also exist other sources of incongruence.

The GGI analysis aided our understanding of whether differences among gene trees are truly due to ILS or to gene tree estimation errors—caused, for example, by low informativeness of UCE loci. It should be noted that few of our loci alone significantly supported either the concatenated (0.02%) or the species tree topology (0.04%), and most importantly, none of these individual gene tree topologies were significantly better than the null hypothesis, where families were placed as a polytomy at the crown node of the Gymnotiformes. Therefore, UCE loci, when taken individually, contain very little phylogenetic signal for resolving deeper nodes. At the same time, it is not surprising to find that the mutation rate of these markers is slower than the speciation rate of this rapid radiation. There were also very few loci supporting each of the 105 hypotheses tested, a pattern that can be attributed to extensive ILS during the early diversification of the Gymnotiformes (Suh et al. 2015; Arcila et al. 2017). In any case, the best gene trees do not need to match the species tree when there are four or more lineages in an unrooted tree (Degnan and Rosenberg 2009). Therefore, the most common gene tree selected across all loci by GGI may not necessarily reveal the true species tree, especially in the presence of high ILS. The reconciliation of these GGI-best gene trees into a species tree recovered the same relationships as in the concatenated analysis. On the other hand, when we only used gene trees that were statistically better than the alternatives (n=55), the species tree we inferred agreed with the coalescent species tree using unconstrained gene trees, except in the lack of resolution for the relationship of *E. electricus* within the Pulseoidea (Supplementary Fig. S16). This result suggests a minimal effect of gene tree estimation error on the relationships inferred by our

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

species tree analysis. Also, the most frequent best tree did not match the GGI-based species tree, and this is another indication that ILS had a strong influence on gene tree heterogeneity—if ILS were not the main cause of incongruence, then the most frequent GGI tree could match the species tree (Arcila et al. 2017).

Our results add to a growing body of work showing that ILS is prevalent in ancient and rapid radiations (Whitfield and Lockhart 2007; McCormack et al. 2013; Oliver 2013; Suh et al. 2015; Hosner et al. 2016; Esselstyn et al. 2017), and suggest that discordance is mostly explained by hemiplasy (i.e., the result of random sorting and fixation of alleles after multiple speciation events) (Avise and Robinson 2008) rather than by the anomaly zone. The theoretical reasoning for this is that the internode length—measured in generations—needs to be at least five times the effective population size for there to be a 95% probability that the gene tree is congruent with the species tree (Nichols 2001). Therefore, effective population sizes that are disproportionally larger than the preceding internal branches, can result in gene trees that are discordant with the species tree more often than gene trees that are concordant (i.e., anomalous gene trees) (Degnan and Rosenberg 2006). If the evolutionary history of the Gymnotiformes falls within the anomaly zone, it would be possible that the observed discordances are also due to frequent anomalous gene trees which are misleading with respect to the true species tree. Some studies have suggested that regions of the Ostariophysan tree of life might be affected by this type of artifact, although not specifically within the Gymnotiformes (Chakrabarty et al. 2017). In contrast, others have concluded that "major otophysan lineages [are] not trapped in the anomaly zone" (Arcila et al. 2017). We also consider unlikely that anomalous gene trees are affecting the base of the Gymnotiformes species tree because in the anomaly zone symmetric anomalous gene trees are more likely to be generated if the species tree is asymmetric. However, in our case, we found the

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

opposite pattern: an asymmetric concatenated tree; i.e., four lineages: (Apteronotidae, (Sternopygidae, (Gymnotidae + Rhamphichthyoidea))), which is more likely to support the anomalous gene tree topology (Kubatko and Degnan 2007; Liu et al. 2009), and a symmetric species tree; i.e., (Apteronotidae + Sternopygidae), (Gymnotidae + Rhamphichthyoidea)). Therefore, this result is incompatible with theory that predicts that for anomalous gene trees to occur with four lineages, the species tree must be asymmetric and the gene tree must be symmetric (Degnan and Rosenberg 2006).

As a final consideration, we must keep in mind that rapid radiations are inherently hard to resolve—in some cases even impossible given current methods—either because they represent hard polytomies or because they might be better represented by a phylogenetic network, rather than by strictly bifurcating relationships (Suh et al. 2015). Therefore, placing a root on the tree can be problematic, and eventually result in misleading inferences, especially when all extant outgroups are distantly related to the ingroup as in the case of the Gymnotiformes, which are estimated to have diverged from the other orders of Characiphysi 100-140 Ma (Betancur-R et al. 2013, 2015, 2017; Hughes et al. 2018). Given the virtual polytomy found at the base of crown-Gymnotiformes, different placements of the root could result in any of the alternative hypotheses proposed in this study.

The difficulty for rooting the gymnotiform tree is manifested by the most recent molecular phylogenetic studies that have provided contrasting, yet highly supported, hypotheses for what is their closest relative: the Siluriformes, which is in agreement with the morphological hypothesis (Fink and Fink 1981; Arcila et al. 2017; Hughes et al. 2018), or Characiformes + Siluriformes (Betancur-R et al. 2015; Chakrabarty et al. 2017). To avoid forcing any of these relationships, we included representative outgroups from all Otophysi, and we rooted the tree

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

with the cyprinid *D. rerio*, which has been unambiguously recovered as the sister group to Characiphysi (Betancur-R et al. 2017; Chakrabarty et al. 2017; Faircloth et al. 2018; Hughes et al. 2018). However, in cases like this, when outgroups are highly divergent, the signal that they provide regarding the root location can be lost, and whether this reflects the history of spurious long branch attraction is unknown (Graham et al. 2002). On one hand, long branch attraction could explain the position of Apteronotidae as the earliest splitting lineage in the concatenated analyses, because when using simple models the long branch of the outgroup may attract the longer branches of the tree (i.e., Apteronotidae) (Felsenstein 1978; Hendy and Penny 1989; Holland et al. 2003). On the other hand, coalescent methods, which are considered to be more robust to artifacts of long branch attraction (Liu et al. 2015b), did not resolve this topology. Unfortunately, the lack of any extant or fossil taxa that could be added to break these long branches precludes us from drawing further conclusions regarding this problem.

Evolutionary History of Electric Organs and Signals

The emergence of electric fishes as models in neural and behavioral sciences has led to a deep understanding of the molecular, anatomical, and physiological diversity of electric organs and signals (Bennett 1971; Kirschbaum and Schwassmann 2008; Finger and Piccolino 2011; Gallant et al. 2014). However, the evolutionary diversity of electric fishes has been little studied and restricted to particular genera or species (Crampton and Albert 2006; Lovejoy et al. 2010; Crampton et al. 2013; Picq et al. 2016; Smith et al. 2016). This lack of study is due, in part, to a paucity of knowledge on signal types and ecology of many species of gymnotiforms, but also to the instability of the deepest relationships among lineages. These issues have precluded a

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

thorough investigation of the origin and evolution of EOs, primarily due to difficulties in controlling for the effects of phylogenetic relatedness in comparative studies.

The complexity of adult electric organs, and the abundance of homoplastic characters, hinders the inference of their evolutionary history. For example, studies based on the ontogeny of electric organs have proposed that a myogenic organ generating monophasic wave-type EODs represents the ancestral state. This claim is supported by the embryological origin and early position of type A (pre)electrocytes—those present in Sternopygidae and Apteronotidae—in between muscle fibers, their morphological similarity to myocytes, and the fact that all gymnotiforms for which information is available produce monophasic EODs during their larval stage. These observations suggest a more recent and independent developmental origin of a neurogenic EO in adult apteronotids (Kirschbaum and Schwassmann 2008). Other hypotheses consider that primitive EOs were myogenic and generated pulse-type EODs, similar to those observed in some genera of catfishes—although these organs are not derived from the same muscle cells and cannot be considered homologous (Crampton and Albert 2006). Another argument in favor of a plesiomorphic pulse-type organ condition is that the ancestral electroreceptors, which were also structurally and physiological similar to those of siluriforms, would have been tuned to detect lower frequency pulses (Stoddard 2002). According to this model, wave-type organs would be more derived and originated in the most recent common ancestor of Sinusoidea (Albert 2001; Crampton and Albert 2006).

Our phylogenomic reconstruction is congruent with either of the hypotheses outlined above, and without additional evidence (e.g., fossils) we are not able to reject either a wave-type or a pulse-type plesiomorphic EO, nor even independent origins (Alves-Gomes et al. 1995; Albert 2001; Crampton and Albert 2006; Kirschbaum and Schwassmann 2008; Arnegard et al.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

2010). Regardless of the plesiomorphic characters of EOs, our results suggest that the rapid and early radiation of gymnotiforms was accompanied by a rapid and early diversification of their electrogenic systems which, given the correlation of EOD diversity and habitats, likely contributed to their successful colonization of almost every lowland freshwater environment in the Neotropics (Crampton and Albert 2006). Hence, these evolutionary and ecological patterns support the role of electrogenesis as a driver generating diversity in the Gymnotiformes, both through speciation and community assembly (Crampton 2011).

Performance of Filtering Strategies

All the filtering strategies compared in our study had a direct and positive effect in reducing the disagreement among data sets. All methods met the main goal of data filtering: reducing gene tree disagreement as a means to decrease noise (non-phylogenetic signal), which ultimately is expected to improve the quality of the data and benefit the inference of the species tree (Philippe et al. 2011; Molloy and Warnow 2018). Filtering also increased alignment length and decreased base compositional bias. Based on this observation, it could be argued that filtering the data solely by locus length or %GC is a simpler approach to increase the ratio of phylogenetic to non-phylogenetic signal. However, the relationship between these measures and higher congruency or nodal support is not straightforward, which indicates that different filtering criteria have additional effects, particularly when accounting for saturation or homoplasy (Kuang et al. 2018).

Our results show that the impact of locus filtering on species tree estimates and consistency differs depending on the criterion and the data set used (Table 2, Fig. 3; Supplementary Table S2, Fig. S7). Additionally, when compared to other studies, evidence

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

suggests that filtering efficacy is highly idiosyncratic to the data and evolutionary scenarios, and corroborates the difficulty of finding a single method that outperforms the others under all circumstances (Doyle et al. 2015; Hosner et al. 2016; Meiklejohn et al. 2016; Blom et al. 2017; Gilbert et al. 2018; Kuang et al. 2018; Molloy and Warnow 2018).

For example, in our study case, selecting loci according to their fit to a molecular clock generally performed worse than all the other strategies. On one hand, this does not agree with other studies that found that this method performs better than selecting genes with highest number of parsimony informative sites (Doyle et al. 2015), but on the other hand, this observation concurs with studies showing that the removal of genes with higher rates of evolution—such as those that deviate from the assumptions of a strict molecular clock—can result in an increase in species tree estimation error because these deleted loci are the ones providing resolution at difficult nodes (Huang and Knowles 2016).

Strategies involving targeted filtering or the number of parsimony informative sites performed well in increasing gene tree congruency, however QIRP, which controls for the effect of homoplasy in the estimation of phylogenetic informativeness, had the greatest effect on increasing the monophyly of each family (except in the Rhamphichthyidae when using the two upper quartiles of genes with most parsimonious informative sites). The most dramatic increase was observed for the Sinusoidea node, for which up to 12% more gene trees recovered this group as monophyletic after filtering the data according to their QIRP values. Furthermore, this translated in the largest improvement on nodal support for this relationship, from BS=58 to BS=88 in the species tree. This result shows the high impact that—apparently small—reductions in gene tree heterogeneity might have in increasing support and consistency of species trees.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

The highest support values were always achieved for the complete taxon set, and for the reduced taxon set support was usually higher when allowing for 25% missing data than with no missing data. It is evident, and well documented elsewhere, that discarding too many loci to reduce the relative amount of missing data can be detrimental, particularly in cases of high ILS (Hosner et al. 2016; Molloy and Warnow 2018). Similarly, an insufficient number of taxa can result in stochastic errors, particularly when the taxon sampling is unbalanced across lineages (Heath et al. 2008; Branstetter et al. 2017). Consequently, effective gene filtering that improves species tree estimation requires finding a balance between data quantity versus data quality (Molloy and Warnow 2018). For low ILS, a few highly accurate gene trees are sufficient to estimate the true species tree (Shekhar et al. 2017). This is evident, for example, in the concatenated analyses, which do not consider gene tree heterogeneity and hence do not account for ILS, and did not show changes in support or topology after filtering despite lower number of loci or taxa. On the other hand, when ILS is high, summary coalescent methods require more data for estimating accurate species trees, and their performance improves with increasing number of genes, despite high disagreement (Shekhar et al. 2017; Molloy and Warnow 2018). Future phylogenomic work involving this group of fishes and their allies (Faircloth et al. 2018) will be valuable to evaluate the extent by which adding data—taxa or genes—improves phylogenetic accuracy.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Overall, targeted filtering proved to work as intended, by improving consistency of gene trees and support of targeted regions over non-targeted regions without selecting or constraining topologies. This is a major advantage of the proposed targeted filtering method, because it does not require prior knowledge or assumptions about the "true" topology of the tree, and can be applied to any time scale, independently of the coalescent time required to achieve monophyly of individual gene genealogies. Furthermore, our results support the idea that a thorough exploration of the data is required prior to estimating species trees, including quantification of ILS and gene tree estimation errors; and that with truly challenging phylogenetic histories, it is important to examine multiple alternative analytical methods (Baurain et al. 2007; Whitfield and Lockhart 2007; Gilbert et al. 2018; Molloy and Warnow 2018). In this context, the utility of the targeted filtering approaches we take adds evidence in favor of the benefits of thorough "gene data interrogation", rather than "gene genealogy interrogation" approaches, to extract the most useful and meaningful phylogenetic signal, and make the best use of the data.

Funding

This work was supported by National Science Foundation (NSF) [DEB-1242260 to B.C.F.; DEB-0614334, DEB-0741450, and DEB-1354511 to J.S.A.; DEB-0916695 and DEB-1354149 to P.C.]. VAT was supported by São Paulo Research Foundation (FAPESP) [2018/20806-3]. Portions of this research (phylogenetic analyses) were conducted with high-performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu).

Acknowledgements

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

We thank our many collaborators in Central and South America for their help in the field, and curators and managers of natural history collections for their service and assistance.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

REFERENCES

Aberer A.J., Kobert K., Stamatakis A. 2014. ExaBayes: Massively parallel Bayesian tree inference for the whole-genome era. Mol. Biol. Evol. 31:2553–2556.

Albert J.S. 2001. Species diversity and phylogenetic systematics of American knifefishes (Gymnotiformes, Teleostei). Misc. Publ. Museum Zool. Univ. Michigan 190:1–127.

Albert J.S., Campos-da-Paz R. 1998. Phylogenetic systematics of Gymnotiformes with diagnoses of 58 clades: a review of available data. In: Malabarba L.R., Reis R.E., Vari R.P., Lucena Z.M.S., Lucena C.A.S., editors. Phylogeny and classification of Neotropical fishes. Porto Alegre: EDIPUCRS. p. 419–446.

Albert J.S., Crampton W.G.R. 2005a. Electroreception and electrogenesis. In: Evans D.H., Claiborne J.B., editors. The physiology of fishes. Boca Raton, FL: CRC Press. p. 431–472.

Albert J.S., Crampton W.G.R. 2005b. Diversity and phylogeny of Neotropical electric fishes (Gymnotiformes). In: Bullock T.H., Hopkins C.D., Poppper A.N., Fay R.R., editors. Electroreception. New York, NY: Springer New York. p. 360–409.

Alves-Gomes J.A., Ortí G., Haygood M., Heiligenberg W., Meyer A. 1995. Phylogenetic analysis of the South-American electric fishes (order Gymnotiformes) and the evolution of their electrogenic system—a synthesis based on morphology, electrophysiology, and mitochondrial sequence data. Mol. Biol. Evol. 12:298–318.

Ané C. 2013. BUCKy users. Available from https://groups.google.com/forum/#!topic/bucky-users/qMsSbTqPSng.

Ané C., Larget B., Baum D.A., Smith S.D., Rokas A. 2007. A Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412–426.

Arcila D., Ortí G., Vari R., Armbruster J.W., Stiassny M.L.J., Ko K.D., Sabaj M.H., Lundberg J.,

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Revell L.J., Betancur-R R. 2017. Genome-wide interrogation advances resolution of recalcitrant groups in the tree of life. Nat. Ecol. Evol. 1:0020.

Arnegard M.E., Zwickl D.J., Lu Y., Zakon H.H., Robinson G.E. 2010. Old gene duplication facilitates origin and diversification of an innovative communication system—twice. Proc. Natl. Acad. Sci. USA 107:22172–22177.

Avise J.C., Robinson T.J. 2008. Hemiplasy: a new term in the lexicon of phylogenetics. Syst. Biol. 57:503–507.

Baurain D., Brinkmann H., Philippe H. 2007. Lack of resolution in the animal phylogeny: closely spaced cladogenesis or undetected systematic errors? Mol. Biol. Evol. 24:6–9.

Bennett M.V.L. 1971. Electric organs. Fish Physiol. 5:347–491.

Betancur-R R., Broughton R.E., Wiley E.O., Carpenter K., López J.A., Li C., Holcroft N.I., Arcila D., Sanciangco M., II J.C.C., Zhang F., Buser T., Campbell M.A., Ballesteros J.A., Roa-Varon A., Willis S., Borden W.C., Rowley T., Reneau P.C., Hough D.J., Lu G., Grande T., Arratia G., Ortí G. 2013. The tree of life and a new classification of bony fishes. PLoS Curr. Tree Life. April 18.

Betancur-R R., Ortí G., Pyron R.A. 2015. Fossil-based comparative analyses reveal ancient marine ancestry erased by extinction in ray-finned fishes. Ecol. Lett. 18:441–450.

Betancur-R R., Wiley E.O., Arratia G., Acero A., Bailly N., Miya M., Lecointre G., Ortí G. 2017. Phylogenetic classification of bony fishes. BMC Evol. Biol. 17:162.

Blom M.P.K., Bragg J.G., Potter S., Moritz C. 2017. Accounting for uncertainty in gene tree estimation: summary-coalescent species tree inference in a challenging radiation of Australian lizards. Syst. Biol. 66:352–366.

Bouckaert R., Heled J. 2014. DensiTree 2: seeing trees through the forest. bioRxiv. 012401:doi:

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

https://doi.org/10.1101/012401.

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

Bravo G.A., Antonelli A., Bacon C.D., Bartoszek K., Blom M., Huynh S., Jones G., Knowles L.L., Lamichhaney S., Marcussen T., Morlon H., Nakhleh L., Oxelman B., Pfeil B., Schliep A., Wahlberg N., Werneck F., Wiedenhoeft J., Willows-Munro S., Edwards S.V. 2018. Embracing heterogeneity: building the Tree of Life and the future of phylogenomics. PeerJ Prepr. 6:e26449v3.

Brinkmann H., van der Giezen M., Zhou Y., Poncelin de Raucourt G., Philippe H. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54:743–757.

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

Chung Y., Ané C. 2011. Comparing two Bayesian methods for gene tree/species tree reconstruction: simulations with incomplete lineage lorting and horizontal gene transfer. Syst. Biol. 60:261–275.

Coates C.W., Cox R.T. 1945. A comparison of length and voltage in the electric eel, *Electrophorus electricus* (Linnaeus). Zootoaxa 30:89–93.

Crampton W.G.R. 2011. An ecological perspective on diversity and distributions. In: Albert J.S., Reis R.E., editors. Historical biogeography of Neotropical freshwater fishes. Berkeley, CA:

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

University of California Press. p. 165–189.

Crampton W.G.R., Albert J.S. 2006. Evolution of electric signal diversity in gymnotiform fishes. In: Ladich F., Collin S.P., Moller P., Kapoor B., editors. Communication in fishes. Enfield, NH: Science Publishers. p. 647–731.

Crampton W.G.R., Rodríguez-Cattáneo A., Lovejoy N.R., Caputi A.A. 2013. Proximate and ultimate causes of signal diversity in the electric fish *Gymnotus*. J. Exp. Biol. 216:2523–2541.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24:332–340.

Dornburg A., Fisk J.N., Tamagnan J., Townsend J.P. 2016. PhyInformR: phylogenetic experimental design and phylogenomic data exploration in R. BMC Evol. Biol. 16:262.

Dornburg A., Townsend J.P., Brooks W., Spriggs E., Eytan R.I., Moore J.A., Wainwright P.C., Lemmon A., Lemmon E.M., Near T.J. 2017a. New insights on the sister lineage of percomorph fishes with an anchored hybrid enrichment dataset. Mol. Phylogenet. Evol. 110:27–38.

Dornburg A., Townsend J.P., Wang Z. 2017b. Maximizing power in phylogenetics and phylogenomics: a perspective illuminated by fungal big data. Adv. Genet. 100:1–47.

Doyle V.P., Young R.E., Naylor G.J.P., Brown J.M. 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64:824–837.

Edwards S. V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA 104:5936–5941.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Elbassiouny A.A., Schott R.K., Waddell J.C., Kolmann M.A., Lehmberg E.S., Van Nynatten A., Crampton W.G.R., Chang B.S.W., Lovejoy N.R. 2016. Mitochondrial genomes of the South American electric knifefishes (Order Gymnotiformes). Mitochondrial DNA B Resour. 1:401–403.

Esselstyn J.A., Oliveros C.H., Swanson M.T., Faircloth B.C. 2017. Investigating difficult nodes in the placental mammal tree with expanded taxon sampling and thousands of ultraconserved elements. Genome Biol. Evol. 9:2308–2321.

Faircloth B.C. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. Bioinformatics 32:786–788.

Faircloth B.C., Alda F., Hoekzema K., Burns M.D., Oliveira C., Albert J.S., Melo B.F., Ochoa L.E., Roxo F.F., Chakrabarty P., Sidlauskas B.L., Alfaro M.E. 2018. A target enrichment bait set for studying relationships among ostariophysan fishes. bioRxiv. 432583: doi: https://doi.org/10.1101/432583.

Faircloth B.C., Chang J., Alfaro M.E. 2012a. TAPIR enables high-throughput estimation and comparison of phylogenetic informativeness using locus-specific substitution models. arXiv preprint. arXiv:1202.1215v1 [q-bio.PE].

Faircloth B.C., McCormack J.E., Crawford N.G., Harvey M.G., Brumfield R.T., Glenn T.C. 2012b. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61:717–726.

Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A phylogenomic perspective on the radiation of ray-finned fishes based upon targeted sequencing of ultraconserved elements (UCEs). PLoS One 8:e65923.

Fara E., Gayet M., Taverne L. 2010. The fossil record of Gonorynchiformes. In: Grande T.,

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Poyato-Ariza F.J., Diogo R., editors. Gonorynchiformes and ostariophysan relationships: a comprehensive review. Enfield, NH: Science Publishers. p. 173–226.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Finger S., Piccolino M. 2011. The shocking history of electric fishes: from ancient epochs to the birth of modern neurophysiology. New York, NY: Oxford University Press.

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

Gallant J.R., Traeger L.L., Volkening J.D., Moffett H., Chen P.-H., Novina C.D., Phillips G.N., Anand R., Wells G.B., Pinch M., Güth R., Unguez G.A., Albert J.S., Zakon H.H., Samanta M.P., Sussman M.R. 2014. Genomic basis for the convergent evolution of electric organs. Science 344:1521–1525.

Gayet M., Meunier F., Kirschbaum F. 1994. Ellisella kirschbaumi Gayet & Meunier, 1991, gymnotiforme fossile de Bolivie et ses relations phylogénétiques au seins des forms actuelles. Cybium 18:273–306.

Gee H. 2003. Evolution: ending incongruence. Nature 425:782.

Gilbert P.S., Chang J., Pan C., Sobel E.M., Sinsheimer J.S., Faircloth B.C., Alfaro M.E. 2015. Genome-wide ultraconserved elements exhibit higher phylogenetic informativeness than traditional gene markers in percomorph fishes. Mol. Phylogenet. Evol. 92:140–146.

Gilbert P.S., Wu J., Simon M.W., Sinsheimer J.S., Alfaro M.E. 2018. Filtering nucleotide sites by phylogenetic signal to noise ratio increases confidence in the Neoaves phylogeny generated from ultraconserved elements. Mol. Phylogenet. Evol. 126:116–128.

Graham S.W., Olmstead R.G., Barrett S.C.H. 2002. Rooting phylogenetic trees with distant

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

outgroups: a case study from the commelinoid monocots. Mol. Biol. Evol. 19:1769–1781.

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

Heled J., Drummond A.J. 2008. Bayesian inference of population size history from multiple loci. BMC Evol. Biol. 8:289.

Hendy M.D., Penny D. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297.

Holland B.R., Penny D., Hendy M.D. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—A simulation study. Syst. Biol. 52:229–238.

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

Huang H., Knowles L.L. 2016. Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. Syst. Biol. 65:357–365.

Hughes L.C., Ortí G., Huang Y., Sun Y., Baldwin C.C., Thompson A.W., Arcila D., Betancur-R R., Li C., Becker L., Bellora N., Zhao X., Li X., Wang M., Fang C., Xie B., Zhou Z., Huang H., Chen S., Venkatesh B., Shi Q. 2018. Comprehensive phylogeny of ray-finned fishes (Actinopterygii) based on transcriptomic and genomic data. Proc. Natl. Acad. Sci. USA 115:6249–6254.

Janzen F.H. 2016. Molecular phylogeny of the Neotropical knifefishes of the order Gymnotiformes (Actinopterygii) (MSc Thesis). University of Toronto: Toronto.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Johansen K., Lenfant C., K S.-N., Petersen J. 1968. Gas exchange and the control of breathing in the electric eel, *Electrophorus electricus*. Z. Vgl. Physiol. 61:137–163.

Kirschbaum F., Schwassmann H. 2008. Ontogeny and evolution of electric organs in gymnotiform fish. J. Physiol. Paris 102:347–356.

Knowles L.L., Lanier H., Klimov P., He Q. 2012. Full modeling versus summarizing gene-tree uncertainty: method choice and species-tree accuracy. Mol. Phylogenet. Evol. 65:501–509.

Kuang T., Tornabene L., Li J., Jiang J., Chakrabarty P., Sparks J.S., Naylor G.J.P., Li C. 2018. Phylogenomic analysis on the exceptionally diverse fish clade Gobioidei (Actinopterygii: Gobiiformes) and data-filtering based on molecular clocklikeness. Mol. Phylogenet. Evol. 128:192–202.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56:17–24.

Kuhner M., Yamato J. 2015. Practical performance of tree comparison metrics. Syst. Biol. 64:205–214.

Lanfear R., Calcott B., Kainer D., Mayer C., Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol. 14:82.

Lanfear R., Frandsen P.B., Wright A.M., Senfeld T., Calcott B. 2016. PartitionFinder 2: New methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772–773.

Larget B.R., Kotha S.K., Dewey C.N., Ané C. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. Bioinformatics. 26:2910–2911.

Lavoué S., Miya M., Arnegard M.E., Sullivan J.P., Hopkins C.D., Nishida M. 2012. Comparable ages for the independent origins of electrogenesis in African and South American weakly

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

electric fishes. PLoS One 7:e36287.

Liu L., Wu S., Yu L. 2015a. Coalescent methods for estimating species trees from phylogenomic data. J. Syst. Evol. 53:380–390.

Liu L., Xi Z., Davis C.C. 2015b. Coalescent methods are robust to the simultaneous effects of long branchs and incomplete lineage sorting. Mol. Biol. Evol. 32:791–805.

Liu L., Yu L. 2010. Phybase: an R package for species tree analysis. Bioinformatics 26:962–963.

Liu L., Yu L., Kubatko L., Pearl D.K., Edwards S. V. 2009. Coalescent methods for estimating phylogenetic trees. Mol. Phylogenet. Evol. 53:320–328.

López-Giráldez F., Townsend J.P. 2011. PhyDesign: an online application for profiling phylogenetic informativeness. BMC Evol. Biol. 11:152.

Lovejoy N.R., Lester K., Crampton W.G.R., Marques F.P.L., Albert J.S. 2010. Phylogeny, biogeography, and electric signal evolution of Neotropical knifefishes of the genus *Gymnotus* (Osteichthyes: Gymnotidae). Mol. Phylogenet. Evol. 54:278–290.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Maley L.E., Marshall C.R. 1998. The coming of age of molecular systematics. Science 279:505–506.

McCormack J.E., Harvey M.G., Faircloth B.C., Crawford N.G., Glenn T.C. 2013. A phylogeny of birds based on over 1,500 loci collected by target enrichment and high-throughput sequencing. PLoS One 8:54848.

Meiklejohn K.A., Faircloth B.C., Glenn T.C., Kimball R.T., Braun E.L. 2016. Analysis of a rapid evolutionary radiation using Ultraconserved Elements: Evidence for a bias in some

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

multispecies coalescent methods. Syst. Biol. 65:612–627.

Mims M.C., Olden J.D., Shattuck Z.R., Poff N.L. 2010. Life history trait diversity of native freshwater fishes in North America. Ecol. Freshw. Fish. 19:390–400.

Mirarab S., Bayzid M.S., Warnow T. 2014. Evaluating summary methods for multilocus species tree estimation in the presence of incomplete lineage sorting. Syst. Biol. 65:366–380.

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

Moller P. 1995. Electric fishes: history and behavior. London: Chapman & Hall.

Molloy E.K., Warnow T. 2018. To include or not to include: the impact of gene filtering on species tree estimation methods. Syst. Biol. 67:285–303.

Nichols R. 2001. Gene trees and species trees are not the same. Trends Ecol. Evol. 7:358–364.

Nozaki H., Iseki M., Hasegawa M., Misawa K., Nakada T., Sasaki N., Watanabe M. 2007. Phylogeny of primary photosynthetic eukaryotes as deduced from slowly evolving nuclear genes. Mol. Biol. Evol. 24:1592–1595.

Oliver J.C. 2013. Microevolutionary processes generate phylogenomic discordance at ancient divergences. Evolution 67:1823–1830.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Paradis E. 2013. Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. Mol. Phylogenet. Evol. 67:436–444.

Paradis E., Claude J., Strimmer K. 2004. ape: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Philippe H., Brinkmann H., Lavrov D. V, Littlewood T., Manuel M., Wörheide G., Baurain D.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.

Philippe H., Delsuc F., Brinkmann H., Lartillot N. 2005. Phylogenomics. Annu. Rev. Ecol. Evol. Syst. 36:541–562.

Picq S., Alda F., Bermingham E., Krahe R. 2016. Drift-driven evolution of electric signals in a Neotropical knifefish. Evolution 70:2134–2144.

Pinheiro J., Bates D., DebRoy S., Sarkar D., Team R.C. 2018. nlme: Linear and nonlinear mixed effects models. Vienna, Austria: R Foundation for Statistical Computing.

Pond S.L.K., Frost S.D.W., Muse S. V. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679.

Pyron R.A., Hendry C.R., Chou V.M., Lemmon E.M., Lemmon A.R., Burbrink F.T. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). Mol. Phylogenet. Evol. 81:221–231.

R Core Team. 2017. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rambaut A., Suchard M., Xie D., Drummond A.J. 2014. Tracer v1.6, Available from http://tree.bio.ed.ac.uk/software/tracer.

Reddy S., Kimball R.T., Pandey A., Hosner P.A., Braun M.J., Hackett S.J., Han K.-L., Harshman J., Huddleston C.J., Kingston S., Marks B.D., Miglia K.J., Moore W.S., Sheldon F.H., Witt C.C., Yuri T., Braun E.L. 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.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Robinson D., Foulds L. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Roch S., Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theor. Popul. Biol. 100:56–62.

Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

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.

Rusinko J., McPartlon M. 2017. Species tree estimation using Neighbor Joining. J. Theor. Biol. 414:5–7.

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

de Santana C.D., Vari R.P., Wosiacki W.B. 2013. The untold story of the caudal skeleton in the electric eel (Ostariophysi: Gymnotiformes: *Electrophorus*). PLoS One 8:e68719.

Schliep K.P. 2011. phangorn: phylogenetic analysis in R. Bioinformatics 27:592–593.

Shekhar S., Roch S., Mirarab S. 2017. Species tree estimation using ASTRAL: how many genes are enough? arXiv. 1704.06831.

Simmons M.P. 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. Mol. Phylogenet. Evol. 80:267–280.

Smith A.R., Proffitt M.R., Ho W.W., Mullaney C.B., Maldonado-Ocampo J.A., Lovejoy N.R., Alves-Gomes J.A., Smith G.T. 2016. Evolution of electric communication signals in the South American ghost knifefishes (Gymnotiformes: Apteronotidae): A phylogenetic comparative study using a sequence-based phylogeny. J. Physiol. 110:302–313.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Springer M.S., Gatesy J. 2016. The gene tree delusion. Mol. Phylogenet. Evol. 94:1–33.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stoddard P.K. 2002. The evolutionary origins of electric signal complexity. J. Physiol. Paris 96:485–491.

Streicher J.W., Schulte II J.A., Wiens J.J. 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empiral study in iguanian lizards. Syst. Biol. 65:128–145.

Su Z., Townsend J.P. 2015. Utility of characters evolving at diverse rates of evolution to resolve quartet trees with unequal branch lengths: analytical predictions of long-branch effects. BMC Evol. Biol. 15:86.

Suh A., Smeds L., Ellegren H. 2015. The dynamics of incomplete lineage sorting across the ancient adaptive radiation of neoavian birds. PLoS Biol. 13:e1002224.

Swofford D.L. 2003. PAUP*. Phylogenetic analysis using parsimony (* and other methods). Version 4. Sunderland, Massachusetts: Sinauer Associates.

Tagliacollo V.A. 2015. Parametric biogeography and the origins of diversity in Neotropical freshwater fishes (PhD Thesis). University of Louisiana at Lafayette: Lafayette, Louisiana.

Tagliacollo V.A., Bernt M.J., Craig J.M., Oliveira C., Albert J.S. 2016. Model-based total evidence phylogeny of Neotropical electric knifefishes (Teleostei, Gymnotiformes). Mol. Phylogenet. Evol. 95:20–33.

Tagliacollo V.A., Lanfear R. 2018. Estimating improved partitioning schemes for Ultraconserved Elements (UCEs). Mol. Biol. Evol. 35:1798–1811.

Tian R., Losilla M., Lu Y., Yang G., Zakon H. 2017. Molecular evolution of globin genes in

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Gymnotiform electric fishes: relation to hypoxia tolerance. BMC Evol. Biol. 17:51.

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

Townsend J.P., Lopez-Giraldez F. 2010. Optimal selection of gene and ingroup taxon sampling for resolving phylogenetic relationships. Syst. Biol. 59:446–457.

Townsend J.P., Su Z., Tekle Y.I. 2012. Phylogenetic signal and noise: predicting the power of a data set to resolve phylogeny. Syst. Biol. 61:835–849.

Triques M. 1993. Filogenia dos gêneros de Gymnotiformes (Actinopterygii, Ostariophysi), com base em caracteres esqueléticos. Comun. do Mus. Ciências da PUCRS, Série Zool. 6:85–130.

Whitfield J.B., Lockhart P.J. 2007. Deciphering ancient rapid radiations. Trends Ecol. Evol. 22:258–265.

Zhong B., Deusch O., Goremykin V.V., Penny D., Biggs P.J., Atherton R.A., Nikiforova S.V., Lockhart P.J. 2011. Systematic error in seed plant phylogenomics. Genome Biol. Evol. 3:1340–1348.

Zimmermann T., Mirarab S., Warnow T. 2014. BBCA: Improving the scalability of *BEAST using random binning. BMC Genomics 15:S11.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

FIGURE CAPTIONS

Figure 1. Graphic outline of all the analyses carried out in this study for each of the UCE data sets of gymnotiforms.

Figure 2. Phylogram of the Maximum Likelihood tree inferred in RAxML using the concatenated full taxon set of Gymnotiformes. All nodes are supported by bootstrap values=100 and Bayesian posterior probabilities=1.00 (from ExaBayes), unless noted.

Figure 3. Species tree inferred in ASTRAL-II for the full taxon set of Gymnotiformes. Black bullets indicate nodes with bootstrap values=100. Bullets in white or shades of gray indicate nodes with low support (support is indicated adjacent to the node) or those nodes not recovered across all analyses (see text). The two sets of squares below nodes represent bootstrap values obtained for the same analysis across all filtered data sets using the reduced taxon set and 25% missing data (to the left) and the reduced taxon set with no missing data (to the right).

Figure 4. Species tree cloudogram inferred in *BEAST for the reduced taxon set and no missing data. The consensus species tree is highlighted in blue. Symbols of electric organ discharges, in purple and gold, indicate the monophyly of families with wave-type (Sinusoidea) and pulse-type (Pulseoidea) electric signals, respectively (colors in the online version of this article). All nodes are supported by Bayesian posterior probabilities=1.00.

Figure 5. Bar plot showing the reduction in Robinson-Foulds distances (Δ%RF) among gene trees after filtering loci across different data sets.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Supplementary Material

Table S1. Concordance factors and their 95% confidence interval for the most frequent bipartitions in the concordance tree inferred from the Bayesian concordance analysis in BUCKy with values of α=1, 5, 10 and ∞.

Table S2. Bootstrap support values recovered for the major nodes of the Gymnotiformes species tree inferred in ASTRAL-II for each one of the filtered and non-filtered data sets. Asterisks (*) indicate incongruence among data sets.

Table S3. Base frequencies and percentage of GC (%GC) content in each major lineage and family across all filtered and unfiltered data sets of the full taxon set. Chi-Square homogeneity tests *P*-values in bold indicate significance at *P*<0.05.

Table S4. Results from the Linear Mixed Effects Models used to test for significant differences in GC composition among filtered and non-filtered data sets.

Suppl. Fig. S1. Summary of previously recovered hypotheses of Gymnotiformes relationships. Hypotheses that recovered monophyletic Sinusoidea or Pulseoidea are indicated with symbols of electric organ discharges in purple and gold, respectively. Asterisks indicate families that are not accepted in the current taxonomy. Fish line drawings by Maxwell J. Bernt.

Suppl. Fig. S2. Maximum Likelihood tree for reduced taxon sets with 25% missing data (a) and no missing data (b).

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Suppl. Fig. S3. ASTRAL-II species tree for reduced taxon sets with 25% missing data (a) and no missing data (b).

Suppl. Fig. S4. Primary concordance and population trees inferred from Bayesian concordance analysis in BUCKy for the non-filtered and filtered data sets. Concordance values are indicated in the trees inferred for the non-filtered data sets using $\alpha=\infty$.

Suppl. Fig. S5. Histograms showing the distribution of gene trees expected simulated under the coalescent model for the inferred species trees (light colors), and of the observed gene trees (dark colors) in each of the analyzed data sets for the full taxon set. Values for the Student's *t*-test and their associated *P*-values are shown in for each plot.

Suppl. Fig. S6. Results plot from the Gene Genealogy Interrogation (GGI) analysis testing 105 alternative hypotheses for the relationships among families of Gymnotiformes (a), and 106 hypotheses (i.e., 105 hypotheses + the null hypothesis of a polytomy at the base of the tree) (b). The *x* axis represents the cumulative number of loci supporting each hypothesis. The green line indicates the topology supported by the concatenated phylogenomic analyses, the orange line represents the topology of the species tree obtained in the coalescent analyses, and the blue line is the null hypothesis. The *y* axis represents associated *P*-values from the approximately unbiased test (AU) for each topology. Only values above the dashed line are hypotheses significantly better than the alternatives.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Suppl. Fig. S7. Summary of phylogenetic hypotheses recovered across all analyses and data sets.

Suppl. Fig. S8. ASTRAL-II species trees inferred for the full taxon set data using the 1st and 2nd quartiles of loci filtered according to the number of parsimony informative sites (a), clock-likeness (b), phylogenetic informativeness (c), and QIRP (d)

Suppl. Fig. S9. ASTRAL-II species trees inferred for the full taxon set data using the 1st quartile of loci filtered according to the number of parsimony informative sites (a), clock-likeness (b), phylogenetic informativeness (c), and QIRP (d)

Suppl. Fig. S10. ASTRAL-II species trees inferred for the reduced taxon set with 25% of missing data using the 1st and 2nd quartiles of loci filtered according to the number of parsimony informative sites (a), clock-likeness (b), phylogenetic informativeness (c), and QIRP (d)

Suppl. Fig. S11. ASTRAL-II species trees inferred for the reduced taxon set with 25% of missing data using the 1st quartile of loci filtered according to the number of parsimony informative sites (b), clock-likeness (b), phylogenetic informativeness (c), and QIRP (d)

Suppl. Fig. S12. ASTRAL-II species trees inferred for the reduced taxon set and no missing data using the 1st and 2nd quartiles of loci filtered according to the number of parsimony informative sites (a), clock-likeness (b), phylogenetic informativeness (c), and QIRP (d)

Suppl. Fig. S13. ASTRAL-II species trees inferred for the reduced taxon set and no missing data

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/sy085/5232148 by Bukkyo University user on 06 December 2018

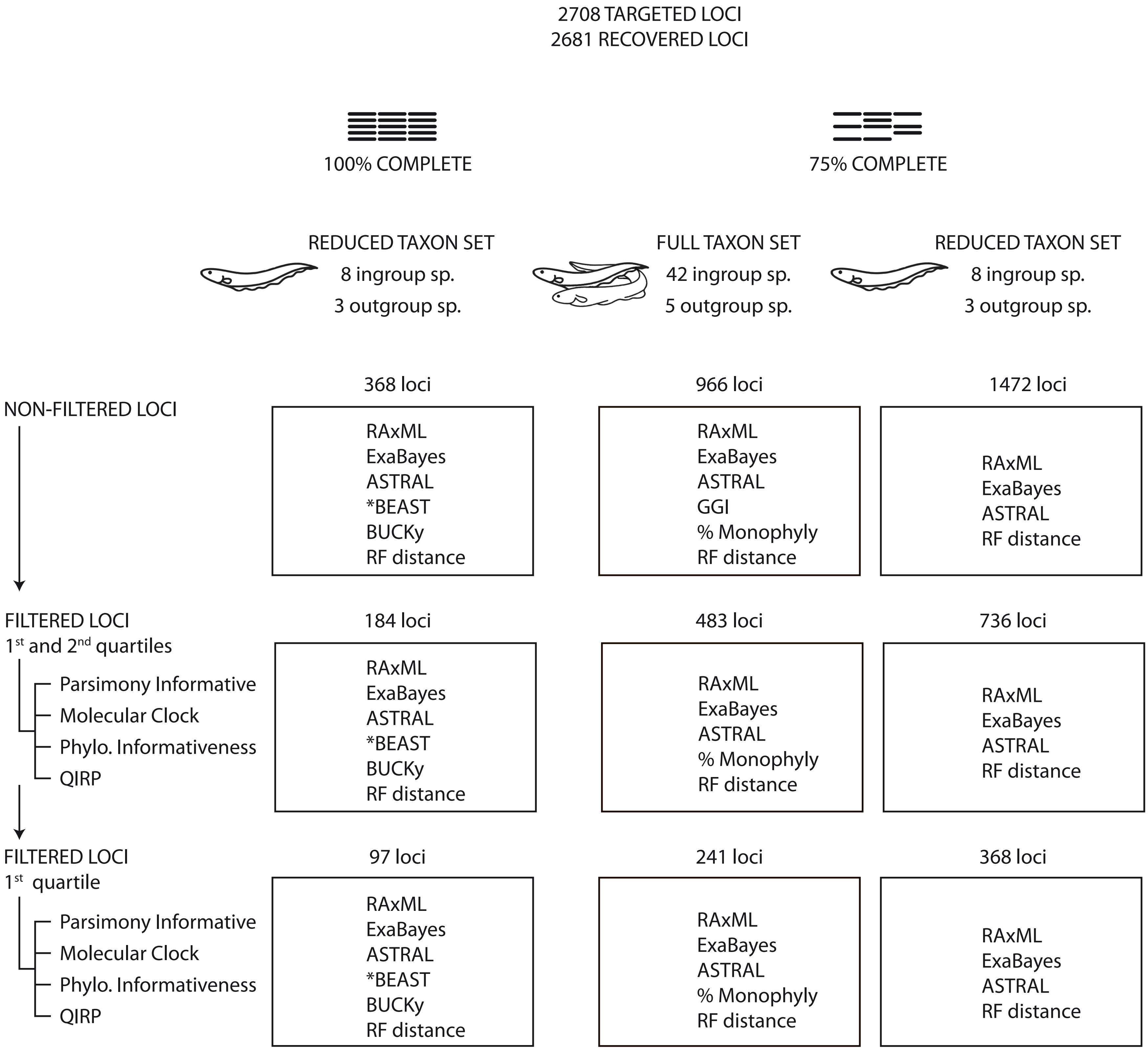
using the 1st quartile of loci filtered according to the number of parsimony informative sites (a), clock-likeness (b), phylogenetic informativeness (c), and QIRP (d)

Suppl. Fig. S14. Scatterplots showing the relationship between locus length and each of the criteria utilized to filter the complete taxon set. Values of Pearson's *r* correlations and associated *P*-values are indicated for each plot. Regression lines are shown in orange.

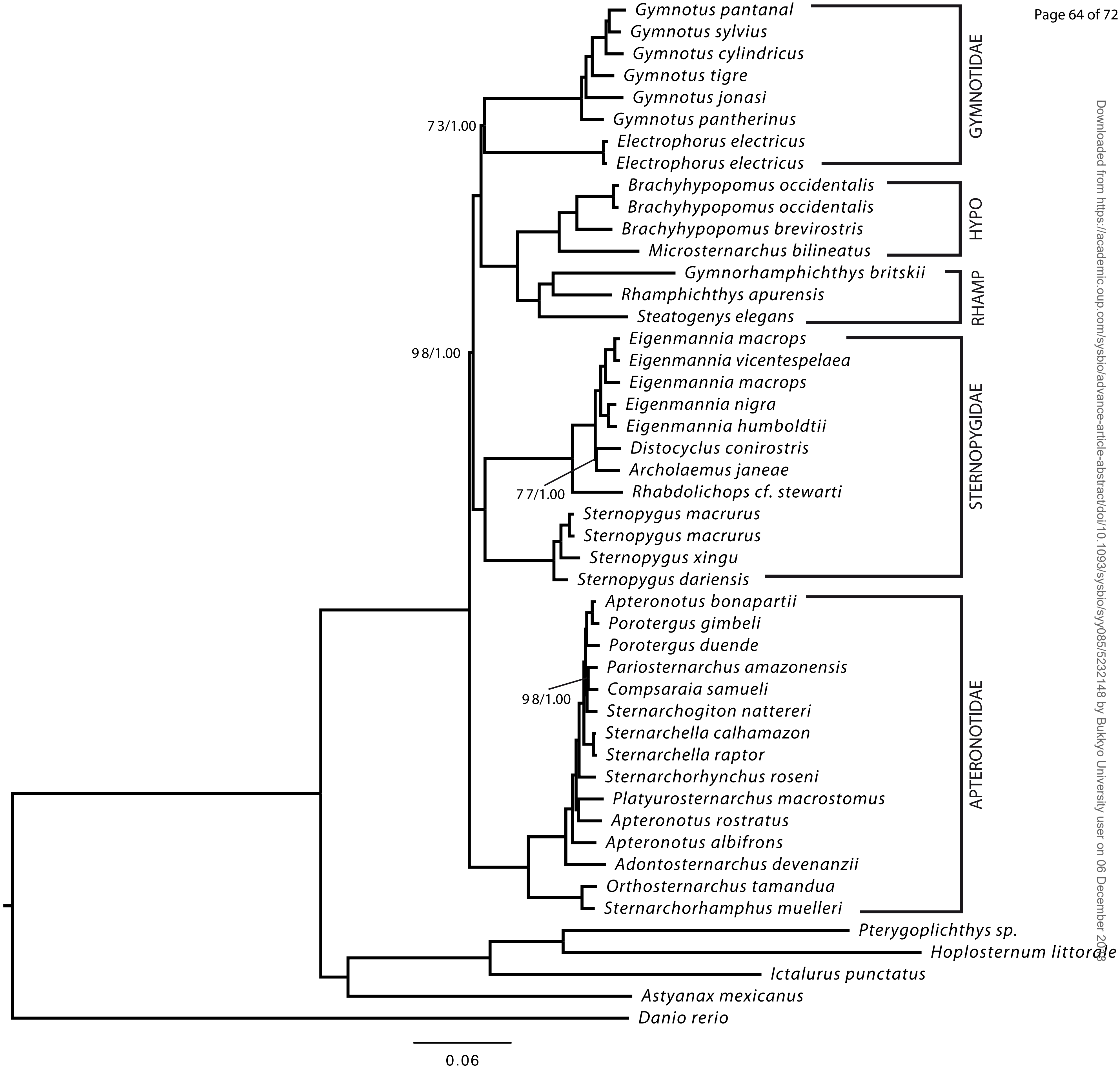
Suppl. Fig. S15. Graphical representation of the reduction in GC content (%GC) for each family of Gymnotiformes following locus filtering of the complete taxon set.

Suppl. Fig. S16. GGI species trees estimated in ASTRAL-II for the best fit trees (a), and for only the best trees that are significantly better than the second-ranked tree (b).

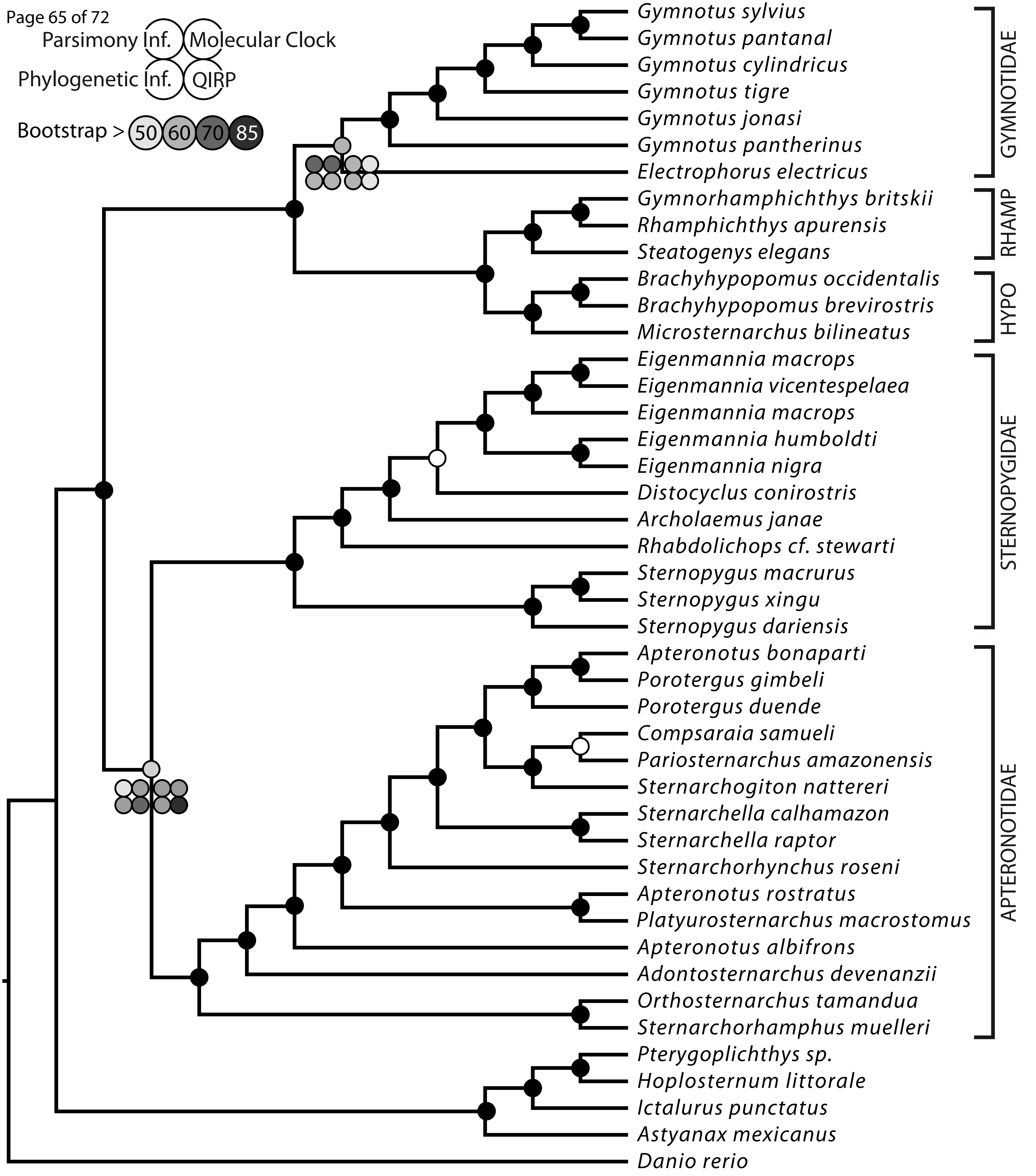
Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018



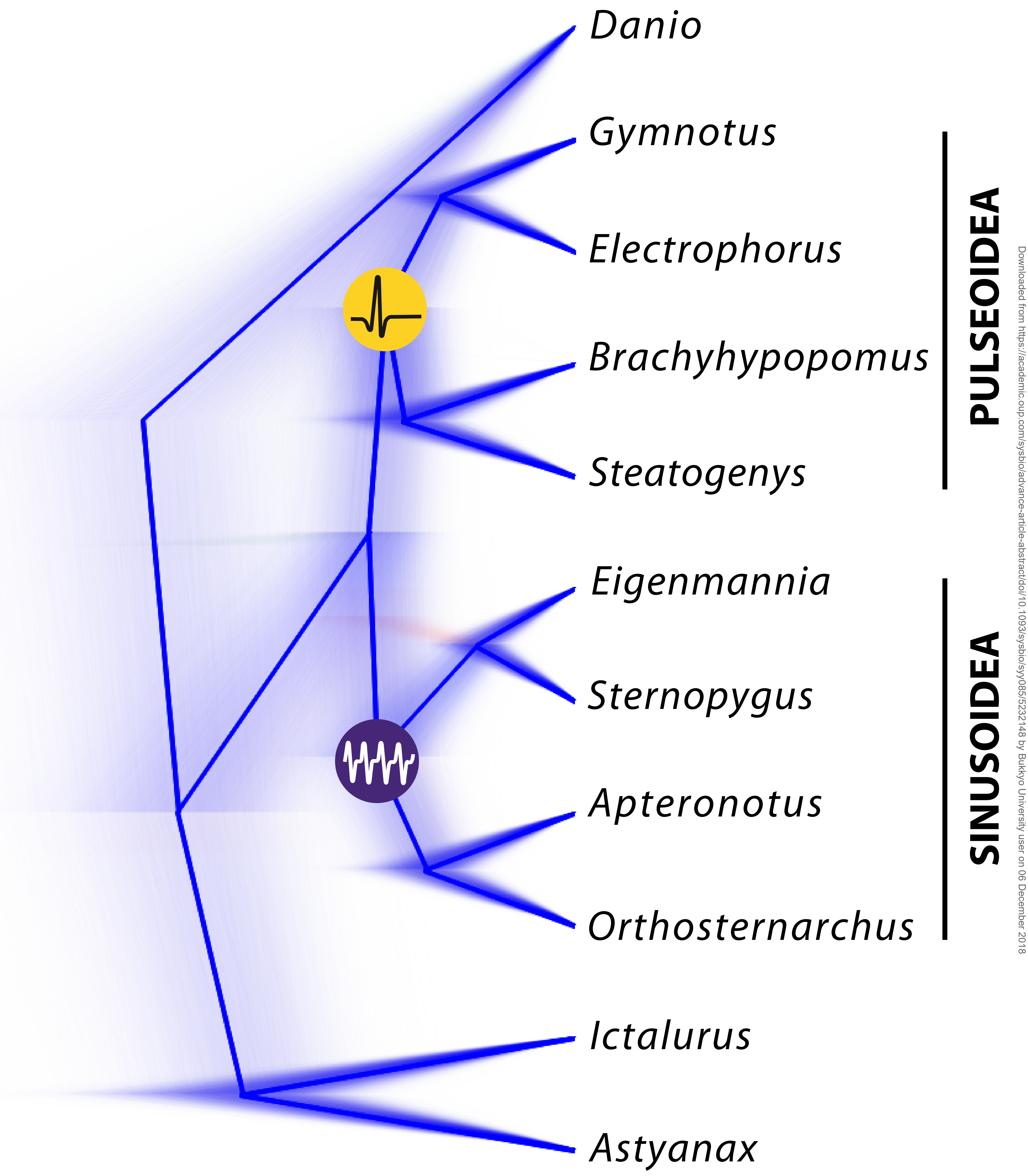
Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018



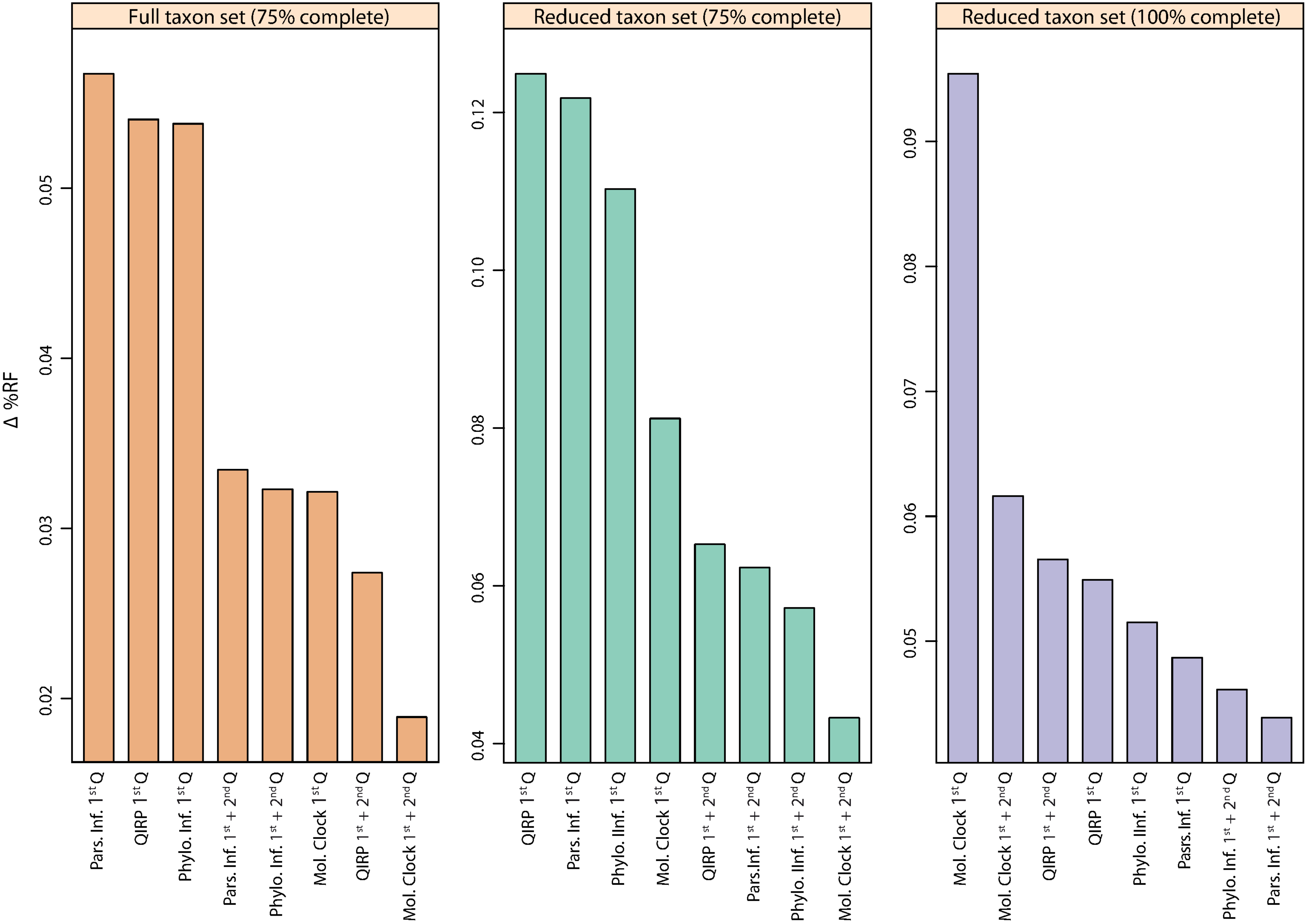
Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018



Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018



Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018



Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

TABLE 1. Family, scientific name, and catalog number of specimen or accession to genome resources, for the species included in this study. Names with an asterisk indicate species included in the reduced taxon set. Abbreviation codes for institution resource collections are provided in the footnote.

Family	Species	Specimen code
Ingroup		
Apteronotidae	*Adontosternarchus devanziaii*	LBP19126
Apteronotidae	*Apteronotus albifrons**	LBP16150
Apteronotidae	*Apteronotus bonaparti*	ANSP200397
Apteronotidae	*Apteronotus rostratus*	stri-26857
Apteronotidae	*Compsaraia samueli*	MUSM54638
Apteronotidae	*Orthosternarchus tamandua**	ANSP200416
Apteronotidae	*Pariosternarchus amazonensis*	ANSP200453
Apteronotidae	*Platyurosternarchus macrostomus*	LBP49302
Apteronotidae	*Porotergus duende*	MUSM54676
Apteronotidae	*Porotergus gimbeli*	ANSP200398
Apteronotidae	*Sternarchella calhamazon*	MCP46987
Apteronotidae	*Sternarchella raptor*	MUSM54640
Apteronotidae	*Sternarchogiton nattereri*	ANSP200449
Apteronotidae	*Sternarchorhamphus muelleri*	ANSP200398
Apteronotidae	*Sternarchorhynchus roseni*	ANSP198343
Gymnotidae	*Electrophorus electricus*	LBP26541
Gymnotidae	*Electrophorus electricus**	Gallant et al. 2014
Gymnotidae	*Gymnotus cylindricus*	LSUMZ1201
Gymnotidae	*Gymnotus jonasi*	LBP34047
Gymnotidae	*Gymnotus pantanal**	LBP32017
Gymnotidae	*Gymnotus pantherinus*	LBP37171
Gymnotidae	*Gymnotus sylvius*	LBP36021
Gymnotidae	*Gymnotus tigre*	JSA060406
Hypopomidae	*Brachyhypopomus brevirostris**	LBP16705
Hypopomidae	*Brachyhypopomus occidentalis*	LSUMZ1849
Hypopomidae	*Brachyhypopomus occidentalis*	stri-203
Hypopomidae	*Microsternarchus bilineatus*	LBP34063
Rhamphichthyidae	*Gymnorhamphichthys britskii*	LBP45898
Rhamphichthyidae	*Rhamphichthys apurensis*	LBP43111
Rhamphichthyidae	*Steatogenys elegans**	ANSP200421
Sternopygidae	*Archolaemus janae*	INPA39971
Sternopygidae	*Distocyclus conirostris*	ANSPt4810
Sternopygidae	*Eigenmannia macrops*	MUSM54627
Sternopygidae	*Eigenmannia macrops*	ANSP198397
Sternopygidae	*Eigenmannia humboldti*	stri-6795
Sternopygidae	*Eigenmannia nigra*	IQ26062014.19
Sternopygidae	*Eigenmannia vicentespelaea**	LBP62040

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

Sternopygidae	*Rhabdolichops cf. stewarti*	LBP41406
Sternopygidae	*Sternopygus dariensis**	stri-14916
Sternopygidae	*Sternopygus macrurus*	LBP16185
Sternopygidae	*Sternopygus macrurus*	LBP46840
Sternopygidae	*Sternopygus xingu*	LBP19643
Outgroup		
Cyprinidae	*Danio rerio**	GCA_000002035.3
Characidae	*Astyanax mexicanus**	GCA_000372685.1
Ictaluridae	*Ictalurus punctatus**	GCA_001660625.1
Loricariidae	*Pterygoplichthys sp.*	LSUMZ5750
Callichthyidae	*Hoplosternum littorale*	LSUMZ6522

ANSP: The Academy of Natural Sciences of Drexel University (Philadelphia, PA).
LBP: Laboratório de Biologia e Genética de Peixes, Departamento de Morfologia, Universidade Estadual Paulista "Júlio de Mesquita Filho", Campus de Botucatu (São Paulo, Brazil).

LSUMZ: Louisiana State University Museum of Natural Science (Baton Rouge, LA).

MCP: Museu de Ciências e Tecnologia, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre (Rio Grande do Sul, Brazil).
MUSM: Museo de Historia Natural at Universidad Nacional Mayor de San Marcos (Lima, Peru).

stri: Smithsonian Tropical Research Institute Neotropical Fish Collection (Panama, Panama).

Other codes refer to uncatalogued specimens.

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018

TABLE 2. Percentage of gene trees recovering the monophyly of each family and major clade of Gymnotiformes for all the filtered and non-filtered datasets analyzed.

		Gymnotiformes	Apteronotidae	Sternopygidae	Rhamphychthyidae	Hypopomidae	Gymnotidae	Sinusoidea	Pulseoidea
Non-Filtered	All loci	84.06	92.75	42.23	70.56	89.42	21.33	14.18	27.43
1st + 2nd quartiles	Parsimony Informative	91.51	97.93	51.35	88.41	95.40	23.81	19.05	34.16
	Molecular Clock	87.31	94.42	43.15	67.34	88.24	19.54	13.45	25.13
	Phylogenetic Inf.	89.65	97.10	50.10	79.41	94.55	23.60	18.84	33.33
	QIRP	91.72	97.31	52.38	81.26	94.57	25.05	21.12	34.16
1st quartile	Parsimony Informative	91.70	98.34	51.04	83.49	91.32	25.31	39.72	39.00
	Molecular Clock	83.76	92.89	38.07	65.09	85.64	17.26	19.29	14.72
	Phylogenetic Inf.	90.04	98.76	52.70	82.27	95.00	25.32	21.99	37.34
	QIRP	91.70	98.76	56.85	83.78	95.82	25.73	26.75	38.17

Downloaded from https://academic.oup.com/sysbio/advance-article-abstract/doi/10.1093/sysbio/syy085/5232148 by Bukkyo University user on 06 December 2018