Molecular Phylogenetics and Evolution xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Phylogenomic analysis on the exceptionally diverse fish clade Gobioidei (Actinopterygii: Gobiiformes) and data-filtering based on molecular clocklikeness

Ting Kuang[a,b,c,1], Luke Tornabene[d,1], Jingyan Li[a,b,c], Jiamei Jiang[a,b,c], Prosanta Chakrabarty[e], John S. Sparks[f], Gavin J.P. Naylor[g], Chenhong Li[a,b,c,*]

[a] Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Shanghai, China
[b] Shanghai Collaborative Innovation for Aquatic Animal Genetics and Breeding, Shanghai, China
[c] National Demonstration Center for Experimental Fisheries Science Education (Shanghai Ocean University), China
[d] School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA 98105, USA
[e] Louisiana State University, Museum of Natural Science, Department of Biological Sciences, Baton Rouge, LA 70803, USA
[f] American Museum of Natural History, Central Park West at 79th Street, NY, NY 10024, USA
[g] University of Florida, Gainesville, FL 3261, USA

ARTICLE INFO

Keywords:
Phylogenomics
Phylogenetics
Gobioidei
Target-gene enrichment
Data filtering
Molecular clocklikeness

ABSTRACT

The use of genome-scale data to infer phylogenetic relationships has gained in popularity in recent years due to the progress made in target-gene capture and sequencing techniques. Data filtering, the approach of excluding data inconsistent with the model from analyses, presumably could alleviate problems caused by systematic errors in phylogenetic inference. Different data filtering criteria, such as those based on evolutionary rate and molecular clocklikeness as well as others have been proposed for selecting useful phylogenetic markers, yet few studies have tested these criteria using phylogenomic data. We developed a novel set of single-copy nuclear coding markers to capture thousands of target genes in gobioid fishes, a species-rich lineages of vertebrates, and tested the effects of data-filtering methods based on substitution rate and molecular clocklikeness while attempting to control for the compounding effects of missing data and variation in locus length. We found that molecular clocklikeness was a better predictor than overall substitution rate for phylogenetic usefulness of molecular markers in our study. In addition, when the 100 best ranked loci for our predictors were concatenated and analyzed using maximum likelihood, or combined in a coalescent-based species-tree analysis, the resulting trees showed a well-resolved topology of Gobioidei that mostly agrees with previous studies. However, trees generated from the 100 least clocklike frequently recovered conflicting, and in some cases clearly erroneous topologies with strong support, thus indicating strong systematic biases in those datasets. Collectively these results suggest that data filtering has the potential improve the performance of phylogenetic inference when using both a concatenation approach as well as methods that rely on input from individual gene trees (i.e. coalescent species-tree approaches), which may be preferred in scenarios where incomplete lineage sorting is likely to be an issue.

1. Introduction

Our ability to reconstruct the Tree of Life has benefited tremendously from the advancement of sequencing technology in recent years (Faircloth et al., 2012; Lemmon et al., 2012; Jarvis et al., 2014). Genome-scale sequence data have become more affordable due to the rapid lowering of costs for high-throughput sequencing. Many researchers are now applying genome-scale data or even full genomes to investigate elusive phylogenetic questions across a wide range of evolutionary time scales (Burleigh et al., 2011; McCormack et al., 2012; Jarvis et al., 2014; Ilves and López–Fernández, 2014; Longo and Bernardi, 2015; Prum et al., 2015). The optimism about reconstructing phylogenies using genome-scale data, dubbed “phylogenomics”, is rooted in the assumption that the problems frequently observed in

* Corresponding author at: Shanghai Universities Key Laboratory of Marine Animal Taxonomy and Evolution, Shanghai, China.
E-mail address: chli@shou.edu.cn (C. Li).
[1] Co-first authors.

https://doi.org/10.1016/j.ympev.2018.07.018
Received 23 August 2017; Received in revised form 11 July 2018; Accepted 17 July 2018

studies using single or a few genes can be overcome by employing large datasets of hundreds of genomic fragments (Gee, 2003). However, incongruence between different analyses still exists despite the use of large-scale datasets (Dunn et al., 2008; Philippe et al., 2009; Schierwater et al., 2009) which suggests that merely adding more sequences is not enough to resolve incongruence but could instead further complicate data analyses (Philippe et al., 2011; Salichos and Rokas, 2013). Thus, there has recently been a concerted effort to understand the phylogenetic utility of phylogenomic datasets with varying sizes, degrees of completeness, and statistical properties (i.e. substitution rates, compositional biases, etc.; Dornburg et al., 2014; Sharma et al., 2014; Eytan et al., 2015; Streicher et al., 2016; Edwards, 2016).

Stochastic errors (sampling) often affect phylogenetic inference due to insufficient data resulting from inadequate taxon representation, too few genes or short sequence lengths. Moreover, systematic distortions can mislead phylogenetic reconstruction as well, such as inaccurately identified homologies (Philippe et al., 2011), erroneous sequence alignment (Lake, 1991; Liu et al., 2010), or inappropriately modeled attributes of the molecular evolutionary process such as base compositional bias (Phillips et al., 2004; Jeffroy et al., 2006), rate variation among different sites, and incomplete lineage sorting (Philippe et al., 2005). While genome-scale datasets can reduce sampling error, they can exacerbate the problem of systematic error if not handled appropriately, leading to incorrect estimations of phylogeny (Phillips et al., 2004; Jeffroy et al., 2006; Rodriguez–Ezpeleta et al., 2007).

To mitigate the influence of these various types of non-phylogenetic signal, phylogeneticists have proposed a wide range of criteria to select genes with strong historical signal for phylogenetic analysis. These include decisive datasets (Dell'Ampio et al., 2014), slow-evolving genes (Nosenko et al., 2013), genes with balanced base composition (Phillips and Penny, 2003), high stemminess value (Fiala and Sokal, 1985; Drovetski, 2002; Qiao et al., 2006), high phylogenetic informativeness (Townsend, 2007; Fong and Fujita, 2011; Lopez–Giraldez et al., 2013), high internode certainty (Salichos and Rokas, 2013), clocklikeness or posterior predictive effect size (Doyle et al., 2015). Additionally, exclusion of rogue taxa and data with ambiguous alignments has been advocated (Rokas and Carroll, 2005; Talavera and Castresana, 2007; Dunn et al., 2008; Philippe et al., 2009). Moreover, missing data may also exacerbate systematic errors (Roure et al., 2013).

Most of these data filtering methods have been developed and tested only on a relatively small number of loci and only recently have some studies tested the effect of individual sources of bias on genome-scale datasets (Salichos and Rokas, 2013; Chen et al., 2015; Doyle et al., 2015; Bossert et al., 2017; Dornburg et al., 2017; Duchêne et al., 2017; Frogoso–Martínez et al., 2017; Qu et al., 2017). With the growing use of genome-scale data in phylogenetic studies, rigorous data-filtering approaches are now logistically feasible and critically important (Phillips et al., 2004).

The fish suborder Gobioidei (Actinopterygii: Teleostei: Gobiiformes) has more than 2200 extant species distributed in marine and freshwaters throughout the tropical, subtropical, and temperate regions of the world. Gobioid fishes represent one of the most evolutionarily successful lineages of vertebrates. The majority of gobioids are benthic or epibenthic fishes, and many possess pelvic fins that are often fused into a “sucker” adapted for life on, or in, the substrate. Most gobioid fishes are small and thus are capable of exploiting a variety of microhabitats in aquatic ecosystems worldwide. In addition, compared to other percomorphs, they exhibit morphological reduction or secondary loss of many of their traits, further complicating morphology-based phylogenetic analyses (Birdsong et al., 1988; Winterbottom, 1993; Akihito et al., 2000; Thacker, 2003; Van Tassell et al., 2011).

Early molecular phylogenies of gobioid fishes were poorly resolved, and in some cases yielded conflicting or incomparable topologies, primarily due to limited taxon sampling, shallow genetic sampling, or differing methods of phylogenetic inference (see Rüber and Agorreta, 2011; Agorreta and Rüber, 2012, for reviews). However, recently

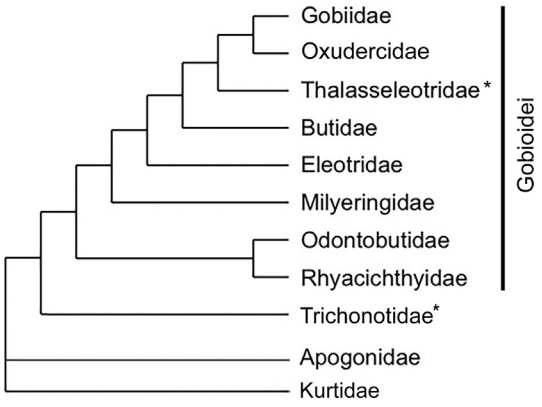

Fig. 1. Current consensus of phylogenetic relationships of Gobioidei. Based on studies using a small number of nuclear genes (Agorreta et al., 2013; Tornabene et al., 2013; Thacker et al., 2015). Asterisks indicate taxa unavailable for this study.

several independent studies featuring increased taxon sampling and larger number of genes, including both mitochondrial DNA and conserved nuclear loci, have converged on a stable backbone phylogeny of Gobioidei (Fig. 1; Thacker, 2009; Thacker and Roje, 2011; Agorreta et al., 2013; Tornabene et al., 2013; Thacker et al., 2015). There are currently-eight families within Gobioidei that are recognized as monophyletic: Rhyacichthyidae, Odontobutidae, Eleotridae, Milyeringidae, Butidae, Thalasseleotridae, Gobiidae, and Oxudercidae. Most of the remaining poorly-resolved relationships within the Gobioidei fall within the most diverse family, the Gobiidae – a clade that underwent abbreviated and rapid periods of speciation early in its evolutionary history (Thacker, 2014). There is also uncertainty about which groups represent the closest relatives to Gobioidei. The most recent hypothesis by Thacker et al., (2015) based on 10 nuclear genes, shows strong support for a sister relationship between Gobioidei and the family Trichonotidae, but the position of the related families Apogonidae and Kurtidae relative to the rest of Gobioidei is unclear (Thacker, 2009; Betancur et al., 2013; Thacker et al., 2015).

In this study, we applied a cross-species target gene capture approach (Li et al., 2013) to collect genome-scale data from 36 representative species from seven of the eight families of the suborder Gobioidei, as well as from seven outgroup lineages (Table S1). We then use phylogenomic data from this group to explore how different data filtering and sub-sampling strategies affect inferences of topology from individual gene-trees, trees from concatenated datasets, and coalescent-based species-tree analyses. Specifically, we tested the usefulness of two data filtering approaches through: (1) screening for slow-evolving genes by comparing their average pairwise distance (p-dist) among different taxa; (2) comparing molecular clocklikeness (MCL) of each gene; clocklikeness being the likelihood ratio between the strict molecular clock model and the relaxed molecular clock model (Kumar and Filipski, 2001; Doyle et al., 2015). Our study provided a robust hypothesis about phylogeny of the Gobioidei and additional information for understanding and mitigating the potential sources of systematic bias in genome-scale phylogenetic analyses.

2. Materials and methods

2.1. Sampling and bait design

Fresh and ethanol-preserved tissues were obtained from 36 species of Gobioidei spanning seven of the eight currently recognized families, and seven outgroup species, *Pterapogon kauderni* and *Sphaeramia orbicularis* (Apogonidae), *Kurtus gulliveri* (Kurtidae), *Ecsenius bicolor* and *Salarias fasciatus* (Blenniidae), *Oreochromis niloticus* (Cichlidae), and *Siniperca chuatsi* (Sinipercidae; Supporting Information Table S1). Total

genomic DNA was extracted from fin or muscle tissue using a Tissue DNA kit (Omega Bio-tek, Norcross, GA, USA) and quantified using a NanoDrop 3300 Fluorospectrometer (Thermo Fisher Scientific, Wilmington, DE, USA). The mtDNA COI gene of each species was amplified and sequenced to double check morphological identifications of voucher specimens. A pair of custom primers were used for PCR and sequencing the COI gene (GobyL, GGCAATCACACGMTGATTYTT, and GobyR, ACAAAGGCAGGYTCTTCRAA). The PCR conditions followed Li and Ortí (2007) but with a modified annealing temperature of 60 °C.

No complete genome sequences of gobies were available when capture probes (baits) were designed, so RNA baits were designed based on the sequences of target regions from several other fish species. Using EvolMarkers (Li et al., 2012), 17,817 conservative single-copy nuclear coding sequence regions (CDS) were found comparing eight model fishes whose genomes are available (*Lepisosteus oculatus*, *Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Tetraodon nigroviridis*, *Anguilla japonica*, *Gadus macrocephalus*, and *Oreochromis niloticus*). Biotinylated RNA baits (MYcroarray, Ann Arbor, Michigan) of 120 bp were synthesized with 2× tiling based on the sequences of *O. niloticus* (3,508,171 bp), which was the most closely related species with an annotated genome to our ingroup taxa at the time of bait design. If the length of the target region was less than 120 bp, the bait sequence was padded with thymine (T) to 120 bp at the 3′ end to facilitate bait synthesis. See Supporting Information (detailed materials and methods and Fig. S1) for detailed methods for library preparation, sequencing, and downstream bioinformatics.

2.2. Test for different data filtering approaches: empirical data

To avoid potential effects due to missing data and to focus on effects due to systematic bias, we only used the 570 loci with sequence data present in all 43 species and less than 5% missing data overall, to test the usefulness of different marker filtering approaches. The following two characteristics used to filter phylogenetic data (referred to hereafter as *predictors*) were evaluated as potential indicators for identifying useful phylogenetic markers: (1) average pairwise distance (p-dist), (2) molecular clocklikeness (MCL). The values of these two predictors were calculated for each of the 570 loci.

It has been demonstrated that nuclear coding loci with moderate evolutionary rates are useful when attempting to resolve phylogenies with deep nodes (Li et al., 2007, 2008; Jones et al., 2012; Lang et al., 2013). Thus, we quantified substitution rate by calculating average pairwise distance using a custom Perl script (Supporting Information Supplementary_data_file_1.txt).

To evaluate the molecular clocklikeness (MCL) of each marker, global likelihood ratio was calculated between a tree with a strict molecular clock enforced against a tree without constraints (Muse and Weir, 1992). Under the null hypothesis, the phylogeny is rooted and the branch lengths are enforced to be clock-like. Under the alternative hypothesis, each branch is allowed to vary independently. The alternative hypothesis invokes s - 2 additional parameters, where s is the number of taxa. The likelihood of each gene tree with and without enforcing a molecular clock was estimated using *PAUP* 4.0b10* (Swofford, 2000). A locus with a small MCL value indicates a clock-like rate of evolution, whereas a large MCL value indicates the locus is less clock-like.

We first checked to see whether each of our predictor values for each gene was correlated with the length of the gene. Preliminary analyses found significant ($p < 0.01$) correlations between the length of each locus and its p-distance and MCL (Supporting Information Table S2). To account for this correlation, we divided these predictor values by the length of each locus. Thus, the effects of these predictors can be interpreted as “effects-per-unit-length”. All references to p-distance and MCL predictor variables hereafter refer to the length-corrected values.

The usefulness of each locus as a phylogenetic marker was then assessed as follows. First, we evaluated the phylogenetic utility of each locus by comparing individual gene trees to a ‘reference tree.’ Ideally the reference tree is the species tree or the ‘true’ topology; however, in empirical studies the true tree topology is unknown. Therefore, as a proxy for the ‘true’ topology, we generated a concatenated maximum likelihood (ML) tree reconstructed from the entire 570 loci (226,979 bp) complete-coverage dataset and used this as a reference tree, as concatenated ML trees have been shown to be an adequate approximation of the true phylogeny in cases where the true topology is not available (Capella–Gutierrez et al., 2014). The concatenated reference tree and single gene trees were estimated using GTRCAT model in RAxMLv7.5.7 (Stamatakis et al., 2005; Stamatakis et al., 2008). Because the CAT approximation is not recommended for analyzing dataset with less than 50 taxa (Stamatakis, 2006), we also compared the results with analysis applying GTRGAMMA model partitioning by three coding positions, which has been shown to be effective at capturing most of the heterogeneity in nuclear sequences of fishes (Li et al., 2008). We also generated concatenated trees using a 14,876 loci dataset (80% missing data), and a 5,954 loci dataset (20% missing data). Finally, for comparison to the concatenated reference tree, we also estimated a coalescent-based species-tree using *ASTRAL-II* (Mirarab and Warnow, 2015; Sayyari and Mirarab, 2016), using un-rooted RAxML gene-trees from the 570 loci.

Individual gene trees and the ML reference tree were compared using Branch Score Distance (Kuhner and Felsenstein, 1994) and Robinson-Foulds Distance (Robinson and Foulds, 1981), which were calculated using treeidst.exe in Phylip v3.695 (Felsenstein, 2005). The Robinson-Foulds Distance only accounts for topological differences, whereas the Branch Score Distance reflects differences in both topology and branch length. Because analyses using the Branch Score Distance and the Robinson-Foulds Distance yielded similar results, so we only report the results based on the Branch Score Distance (hereafter “tree distance”). We then used independent and multiple-regression models to test the correlation between the tree distances and the values of each of the two predictors. A strong correlation between the tree distance and the value of a particular predictor, or combination of predictors, suggests that this data-filtering characteristic could be a good indicator of phylogenetic performance and thus a useful tool for data filtering. We performed a Ramsey Regression Equation Specification Error Test (Ramsey, 1969) to determine if non-linear combinations of our predictor values have power to explain tree distance, and thus a linear model would be mis-specified.

We then set out to determine whether combining many genes with poor predictor values would dampen or amplify the effects of systematic biases. We ranked the 570 genes according their values of our two predictors, then took 100 loci with lowest predictor values and 100 loci with highest predictor values, concatenated them respectively, and analyzed them using RAxML v.7.5.7 (GTRCAT model, 500 replicates) (Stamatakis et al., 2005; Stamatakis et al., 2008). We also analyzed these same 100 best/worst loci for each predictor using *ASTRAL-II*. We then compared the topology and support values of the best/worst 100-loci trees to each other and our reference tree to evaluate their performance.

3. Results

3.1. Sequencing reads and assembly

One lane of the Illumina HiSeq2000 run generated a total of 410 million raw reads (average of 9.8 million reads per sample), with *Stonogobiops nematodes* having the most reads (14.7 million) and *Typhleotris pauliani* having the fewest (5 million; Supporting Information, Table S1). After adapter trimming and quality filtering, an average of 9.7 million reads (99.5% of the raw reads) per sample were retained. The reads were then parsed into the different files according to their matching score to the bait sequences of the target loci. An average of 10,397 target loci per sample contained parsed reads. After

de novo assembly using Trinity assembler, on average, 5070 loci had one contig; 5327 loci had more than one contig, which were primarily short fragments that did not overlap enough to be assembled using Trinity's de Bruijn graph approach, or non-target sequences (e.g. paralogues) that were removed later in a reciprocal BLAST step. An average of 463 out of these 5327 loci could be successfully merged into one contig using Geneious (Supporting Information, Table S1). An average of 9795 orthologous sequences (55% of total number of targets) was identified for each sample through pairwise alignments and BLAST searches. The species with the most number of loci captured was *K. gulliveri* (15,003 loci, 84.2% of the total number of targets), while the one had the worst result was *Ctenogobius boleosoma* (6099 loci, 34.2% of the total number of targets). An average of 1.5 million reads (15.1% of the filtered reads) per sample was mapped onto the assembled contigs, with a mean depth of 67× (Supporting Information, Table S1). On average, 50.1% of the 3.5 million bp targeted nucleotide sites were covered for each species. The number of captured loci is negatively correlated with the distance between baits and captured sequences ($p < 0.01$), which ranged from 10.3% to 22.8%, while the number of raw reads could not predict the number of loci obtained ($p = 0.97$) (Supporting Information Fig. S2). The 100 loci with the largest average pairwise distances were examined by eye to identify potential misalignment, of which twenty-four alignments were corrected by hand. No other cases of misalignment were found when 100 additional loci were randomly spot-checked. There were 694 loci with sequence data present for all taxa, of which 570 had less than 5% missing data and were thus used in downstream analyses.

3.2. Gobioid reference tree

The ML analysis on concatenated sequence of 570 loci present in all taxa with less than 5% missing data, using GTRCAT model and GTRGAMMAR model resulted in the same topology, with 100 bootstrap support for all but one node (Fig. 2). We also generated concatenated trees using our complete dataset, a 14,876 loci dataset (80% missing data), and a 5954 loci dataset (20% missing data), all of which resulted in an identical topology to the 570 loci dataset but with all nodes having 100 bootstrap support. The ASTRAL tree generated gene trees from the same 570 loci (Supporting Information Fig. S3) is identical to the ML concatenated tree with the exception of some relationships within the Gobiidae, which were not very well supported in the ASTRAL tree. All families within the Gobioidei were recovered as monophyletic.

3.3. Assessing two methods of data filtering for selecting phylogenetic markers

The two predictors, p-distance and MCL were calculated for each of the 570 loci that were present for all taxa and had less than 5% missing data. The tree distance between the gene tree of each locus and the reference tree was calculated. The raw values of the two predictors and the tree distance for each locus are available as Supporting Information (Supplementary_data_file_2.txt).

We assessed how well each of our two predictors correlated with the tree distance between each gene tree and the reference tree through regression analyses using combination the two predictors, and each the predictors individually. Because both our predictors were significantly correlated with locus length, we divided them by locus length to mitigate the effects of the quantity of data in each locus and focus on systematic sources of bias. After accounting for locus length in these predictors, both predictors were significantly correlated with the tree distance (Fig. 3; Table 1), indicating that the both predictors have some utility in indicating phylogenetic performance. The explanatory power of each of our predictors individually ranged from an R^2 of 0.08 (for p-distance) and 0.14 (for MCL) (Fig. 3; Table 1). The multiple-predictor model passed the Ramsey test (Ramsey, 1969), indicating that non-linear combinations of our data did not explain tree-distance significantly better than a linear model. The multiple-predictor model had an R^2 of 0.15 (Table 1). Although we were not explicitly interested in the effect of data quantity (e.g. locus length) on loci performance, for the purpose of data exploration, we also created two other a multiple regression models that included length as a main effect, as well as "length × predictor" interaction terms (Supporting Information Table S3). This model was created using our raw predictor values (not corrected for length). Because the distribution of the model residuals maybe is non-normal (and highly skewed), we also log-transformed predictor values for regression analyses. These models had higher R^2 values than those in Table 1, as we expected with the addition of locus length as a predictor, and as expected from our preliminary regressions, MCL and p-distance had significant interaction terms with locus-length (Supporting Information Table S3).

We created both concatenated maximum likelihood trees and ASTRAL species-trees from the 100 genes with the best/worst values for p-distance, and MCL, and compared them to each other and our reference tree to evaluate whether systematic biases are dampened or amplified when many genes are used (Figs. 4 and 5). For all predictors, the ASTRAL and concatenated trees from the 100 best ranked loci produced trees that were nearly identical to our reference tree (Supporting Information Figs. S4 and S5). The only exceptions were that relationships within the Gobiidae were variable and generally not well-resolved. However, when looking trees generated from the 100 worst-ranked genes for each predictor, both concatenated trees (trees from genes with the 100 worst MCL, p-distance values; Figs. 4A, 5A), and ASTRAL trees (Figs. 4B, 5B) all had topologies that conflicted with the reference tree, all with strong support. In most cases the conflicting topologies were restricted to the errant placement of one or two branches (e.g. Figs. 4, 5B). However, the concatenated tree from the 100 least clocklike loci had a highly divergent topology, with *Odontamblyopus* and *Ctenogobius* being recovered in a grade outside of all other gobioids (rather than within Oxudercidae), and with Butidae, Eleotridae, Milyeringidae, Rhyacichtyidae, and Odontobutidae all being resolved in one large clade sister to Oxudercidae + Gobiidae (Fig. 5A).

4. Discussion

4.1. Enrichment efficacy

The efficacy of target enrichment protocols may vary depending on bait design and the molecular properties (i.e. GC content) of the targeted regions (Tewhey et al., 2009; Hedges et al., 2011). A more common cause of low capture efficacy is genetic divergence between bait sequences and targets (Hedtke et al., 2013; Li et al., 2013; Bragg et al., 2016; Hugall et al., 2016). We also found that the number of loci captured was highly correlated with the genetic distance between the baits and the target sequences ($R^2 = 0.63$, $p < 0.01$; Fig. S2). We did not find a correlation between the number of loci captured and the amount of raw reads sequenced (Fig. S2), that is more data did not result in more loci captured, which suggests that the amount of data we collected (9.8 million reads per sample on average) was sufficient to obtain the capturable target sequences under the current experimental conditions.

4.2. Phylogenomic hypothesis of Gobioidei

Our results based on the first phylogenomic dataset of the Gobioidei (Fig. 2) strongly support the monophyly of the Gobioidei and of the families within it, and is in agreement with topologies recovered by other recent molecular studies (Fig. 1; Thacker, 2009; Betancur et al., 2013; Thacker et al., 2015). Several relationships recovered by our analysis conflict with results from early molecular phylogenies based on a limited number of genes, including those based entirely on mitochondrial DNA (Thacker and Hardman, 2005; Thacker, 2009;

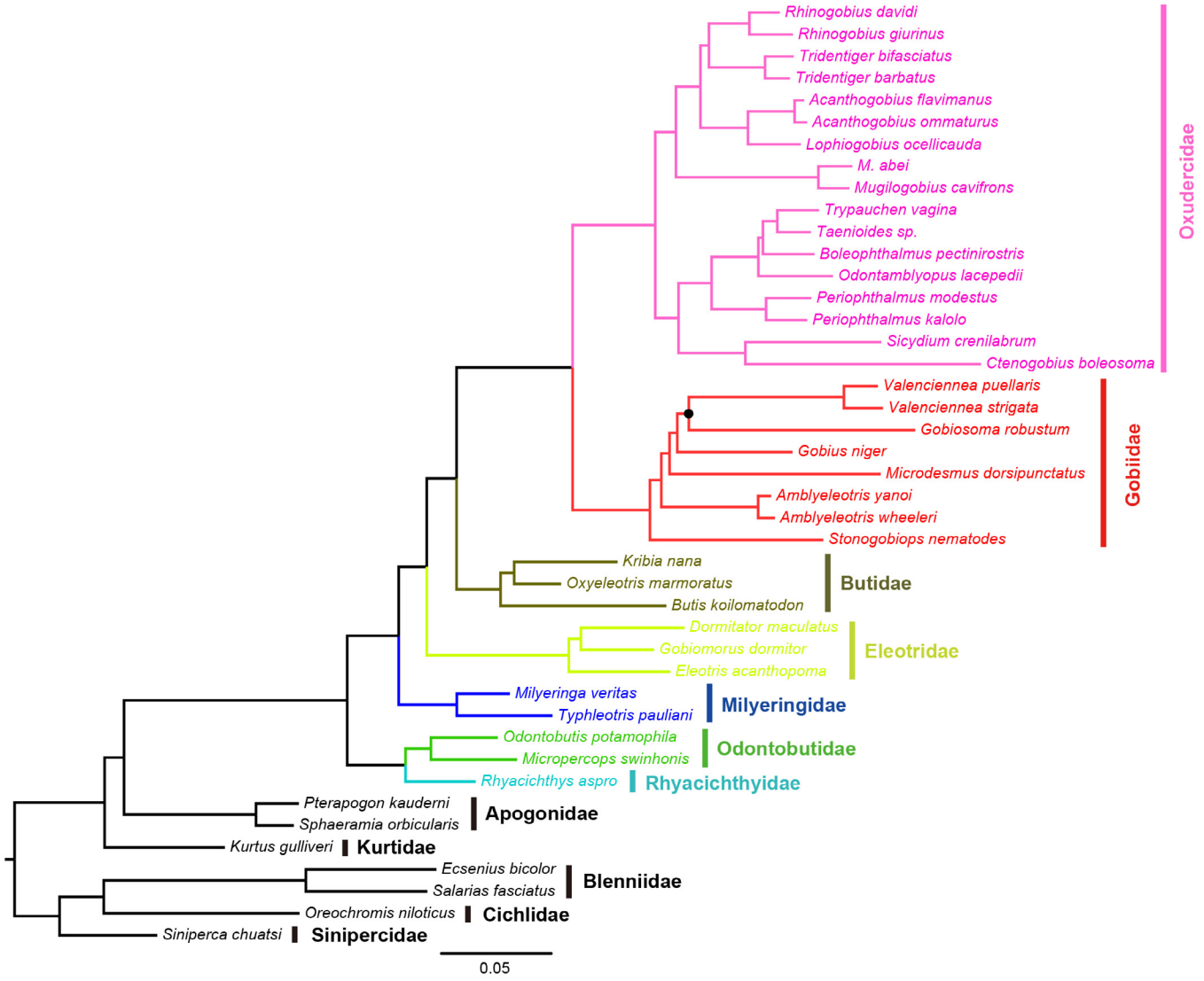

Fig. 2. ML tree based on the complete dataset, 570 loci. No missing taxa for all loci, hereafter referred to as the 'reference tree'. The dot denotes the only node with less than 100% bootstrap support.

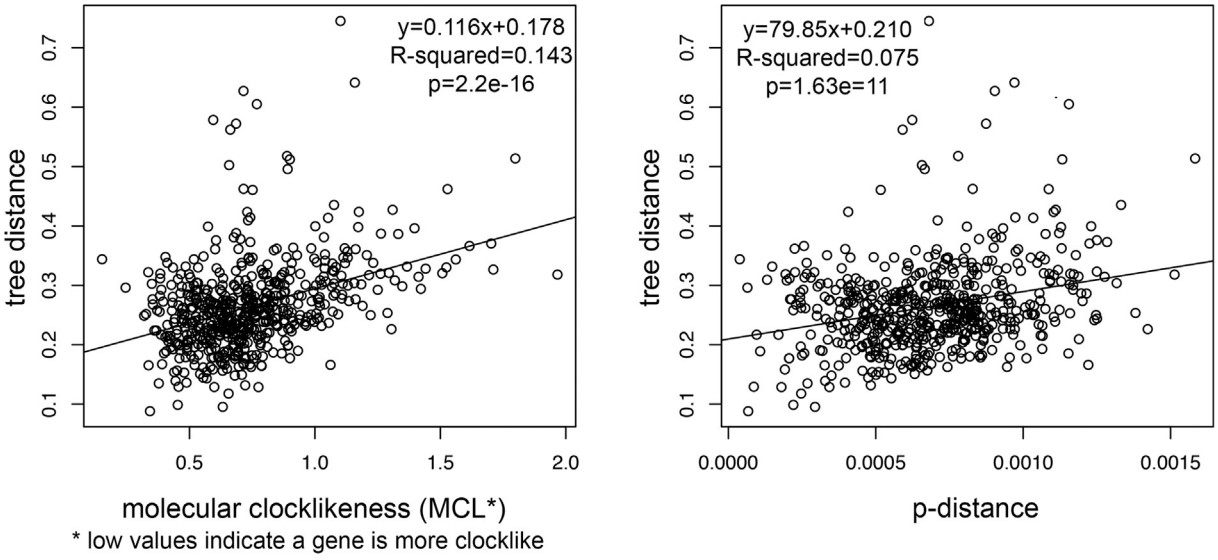

Fig. 3. The relationships between tree distance and values for the two predictors. Values for both predictors have been divided by length.

Table 1
Summary of linear regression models using the two predictors in combination (top) and independently (bottom) to explain tree distance between gene trees and the reference tree. Both predictors have been divided by locus length.

Predictors in model	Regression coefficents		Model Statistics		logLik	AIC
	MCL	p-distance	R2	df		
Multiple predictor model						
MCL, p-distance	0.098	32.864	0.15	4	708.74	−1409.48
Single predictor models (ranked by AIC)						
MCL	0.116	–	0.14	3	705.49	−1404.98
p-distance	–	79.852	0.08	3	683.68	−1361.36

Chakrabarty et al., 2012). For example, the analysis by Thacker and Hardman (2005) suggested that a clade composing Milyeringidae and Odontobutidae was sister group to the remainder of Gobioidei excluding the Rhyacichthyidae, and Thacker (2009) recovered a clade containing Rhyacichthyidae, Odontobutidae and Milyeringidae (represented solely by *Milyeringa*), whose intrarelationships were unresolved, which was the sister group to all other lineages of Gobioidei with strong support. However, the relationships among the major gobioid families here (Fig. 2) are in agreement with more recent studies that have incorporated nuclear genes (Agorreta et al., 2013; Tornabene et al., 2013; Thacker et al., 2015), supporting the importance of choosing appropriate markers for the phylogenetic question at hand. The relationships within the goby family Oxudercidae, although based on very limited taxon sampling, are also congruent with previous studies containing many more species (Agorreta et al., 2013; Tornabene et al., 2013; Thacker, 2013). Relationships within the Gobiidae, while well supported in our reference tree (but poorly supported in all other 100-loci trees discussed below), are based on too few taxa to be critically compared to previous studies with more extensive sample, but in general the relationships shown in Fig. 2 are consistent with those from recent studies (Thacker and Roje, 2011; Agorreta et al., 2013; Tornabene et al., 2013).

The general congruence between our tree based on our large genomic datasets (570 to 14,876 loci) and those containing far fewer loci (e.g. Agorreta et al., 2013; Thacker et al., 2015), suggest that in the case of the backbone phylogeny of Gobioidei, the five to 10 genes (in concert with robust taxon sampling) used in past studies may have been sufficient data to recover the true phylogeny. This may indeed be the case for other studies as well, but this is entirely dependent on the characteristics of the genes chosen. In the case of gobioids, below we show that increasing from 10 genes (Thacker et al., 2015) to 100 genes can actually decrease phylogenetic performance and lead to strongly-supported incorrect topologies if the loci are sources of systematic biases, thus stressing the importance of assessing the quality of data in large-scale phylogenomic analyses.

4.3. Criteria for selecting best molecular markers

To avoid incorrect phylogenetic inference due to potential systematic errors, two approaches could be used. The first approach is to explicitly model all forces that have driven the evolution of the molecules, many of which are often unknown. The second approach is to identify and exclude data that are not following standard time-homogeneous i.i.d. model (Philippe et al., 2005). The increasing use of genome-scale phylogenetic datasets makes the data filtering methods more practical and imperative than ever before, enabling more effective approaches for detecting phylogenetic signal (Phillips et al., 2004). Some of the proposed criteria for choosing/excluding phylogenetic markers include substitution rate (Nosenko et al., 2013), nucleotide base composition (Phillips and Penny, 2003), phylogenetic informativeness (Townsend, 2007; Lopez–Giraldez et al., 2013), clocklikeness or posterior predictive effect size (Doyle et al., 2015), among others.

The results from our gene-capture Gobioidei dataset show that both p-distance and molecular clocklikeness are significantly correlated with the distance between the gene tree and the reference tree. We also note that both of our predictors were strongly correlated with the length of locus (Supporting Information Table S2), that is, longer loci tend to have more favorable characteristics in terms of their substitution rate and clocklikeness. Thus, selecting longer loci may be an effective first-step screening strategy in selecting loci with desirable properties. Indeed, Camargo et al. (2012) found that short loci (~150 bp) were substantially less accurate than 295–440 bp loci, however their study focused on a more recently-divergent group which likely has significantly fewer informative sites per loci on average. However, the relationship between locus length and phylogenetic accuracy is more complex than simply adding more variable sites, and adding more data by combining loci will not unequivocally improve accuracy of analyses if loci still possess sources of systematic bias, as indicated by trees shown in Figs. 4 and 5.

Despite a significant correlation between both of our predictors and the distance between gene trees and the reference tree, the explanatory power of p-distance is relatively low in our dataset, that after accounting for gene length, slow evolving genes were significantly more accurate than faster evolving genes, but this only explained 7.5% of the variation in tree distance (Table 1). Despite the low explanatory power of a genes substitution rate on a locus-by-locus basis, the benefit of filtering out rapidly evolving genes is more pronounced when combining loci. When we concatenated 100 rapidly-evolving genes (Fig. 4A) or analyzed them using ASTRAL (Fig. 4B), we recovered a topology that conflicted with our reference tree, and one that has never been shown in previous studies; *Sicydium* and *Ctenogobius* were recovered sister to the rest of the Oxudercidae with strong support. Both *Sicydium* and *Ctenogobius* belong to the "*Stenogobius* lineage" of Agorreta et al. (2013), and species in this group have historically and consistently been resolved as the sister group to a clade of eel-gobies and mudskippers (represented in our study by *Taenioides*, *Trypauchen*, *Boleopthalmus*, *Odontamblyopus* and *Periophthalmus*) in molecular studies using different datasets (Thacker, 2009, 2013; Agorreta et al., 2013; Tornabene et al., 2013). Like our dataset focusing on 100 rapidly evolving loci (Fig. 4), the datasets from Thacker (2009) and Thacker (2013) dataset consisted only rapidly-evolving mitochondrial genes, ableit with much denser taxon sampling in the Oxudercidae than our study. Nevertheless, Thacker (2009, 2013) recovered a topology consistent with our reference tree, and different than the tree we recover from the 100 loci with the highest evolutionary rates (Fig. 4). This suggests that homoplasy in rapidly evolving genes will obscure phylogenetic signal more so in cases where taxon sampling is sparse and terminal branches are longer, as is the case in our dataset versus those of Thacker (2009; 2013). Our study focuses primarily on organisms that have diverged within the last 60–100 million years (Thacker, 2014), and thus selecting genes with high signal-to-noise ratio may be even more pronounced in studies covering broader evolutionary time-scales where substitution saturation may be more prevalent.

Our data strongly suggested that MCL was the most useful data-filtering tool for our phylogenomic dataset. There was a significant correlation between how clocklike a gene was and how well that gene tree matched the 'true' reference tree (Fig. 3; Table 1). The explanatory power of MCL ($R^2 = 0.14$) was substantially higher than rate. The relationship between clocklikeness and phylogenetic performance held true, and was exacerbated when many non-clocklike loci were analyzed together (Fig. 5). Combining multiple non-clocklike loci in the ASTRAL species-tree analysis yielded a tree that conflicts (with strong support) with our reference tree regarding the placement of *Ctenogobius* (Fig. 5B). Concatenating non-clocklike loci resulted in the worst topology in this study (Fig. 5A), which shows well-supported place of most of the 'basal' gobioid families in a distinct clade sister to Gobiidae + Oxudercidae, and the genera *Odontamblyopus* and *Ctenogobius*

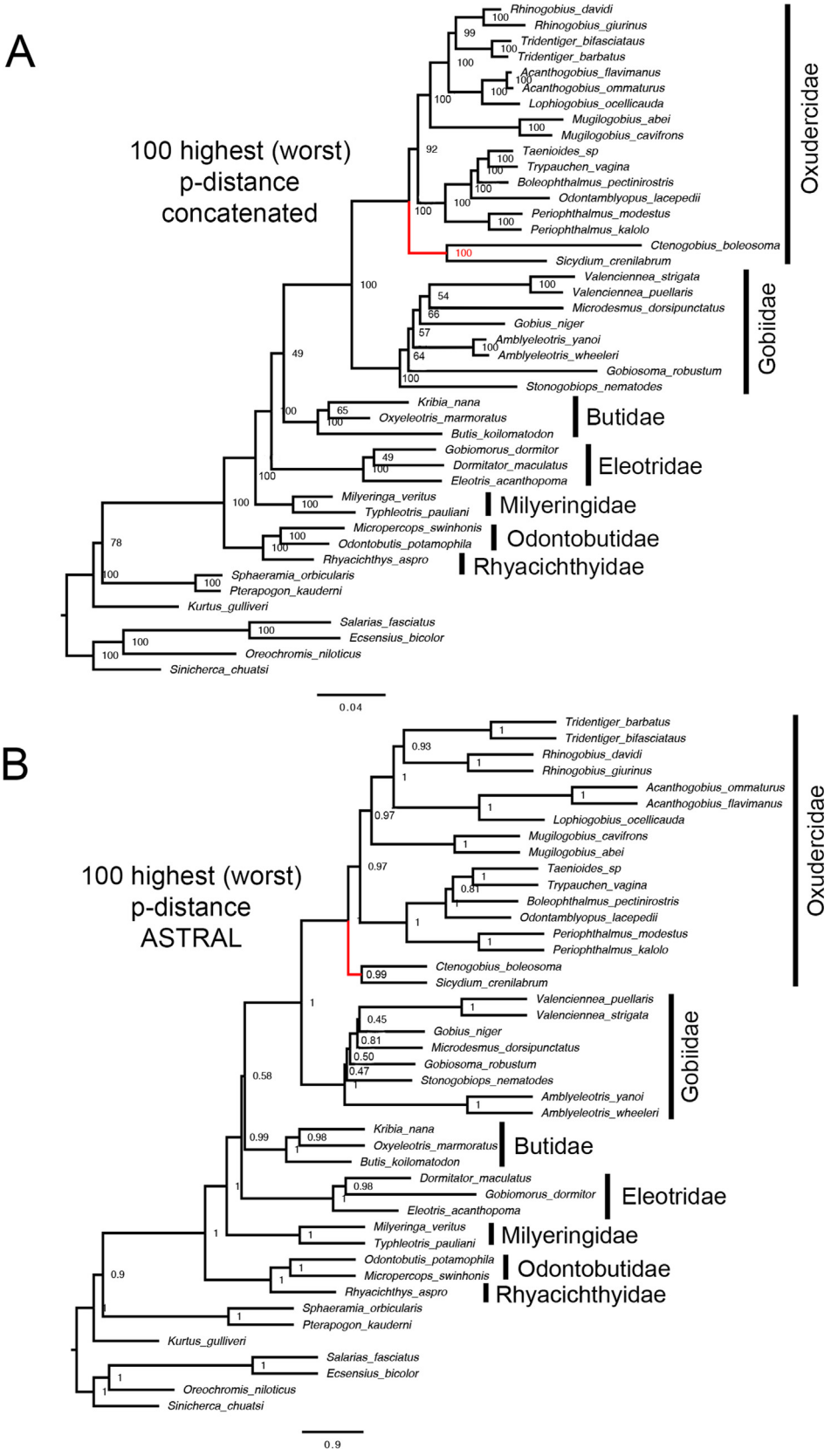

Fig. 4. Trees inferred from 100 loci with highest average pairwise genetic distance (p-distance). (A) concatenated RAxML analysis, support at nodes are bootstrap values. (B) ASTRAL species-tree analysis, support at nodes are local posterior probabilities. Major topological differences from the reference tree are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

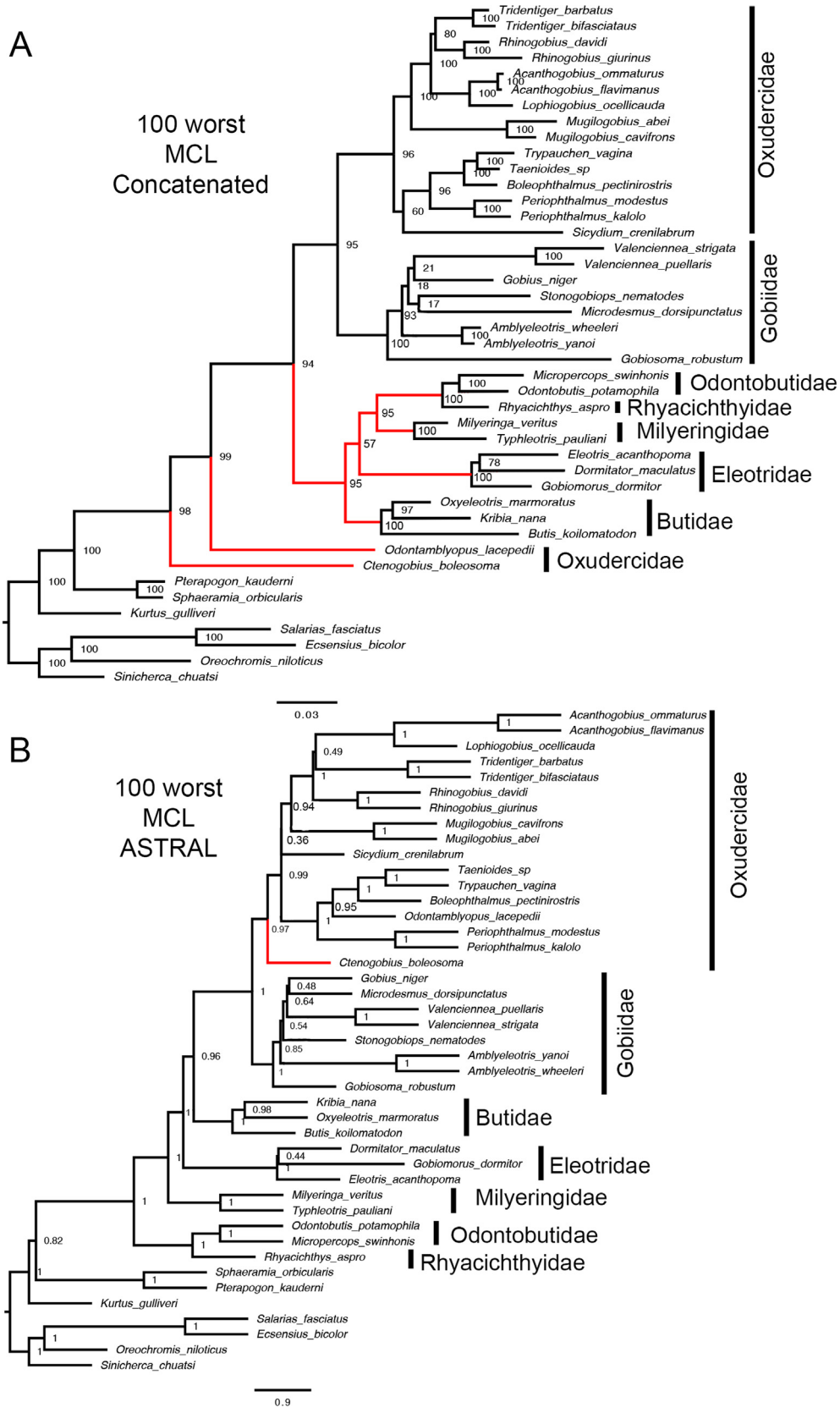

Fig. 5. Trees inferred from 100 least clocklike loci. (A) concatenated RAxML analysis, support at nodes are bootstrap values. (B) ASTRAL species-tree analysis, support at nodes are local posterior probabilities. Major topological differences from the reference tree are highlighted in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

being resolved well-outside the rest of the Gobioidei. These relationships are not supported by any phylogenetic study to date. Doyle et al. (2015) also discovered that clocklikeness could be a good indicator of phylogenetic reliability of genes in yeast and amniotes. Our results present the second study testing clocklikeness as a predictor of phylogenetic signal, which used a novel dataset in a much younger divergence timeframe than that of yeast and amniotes even though more datasets should be examined before generalizing the conclusion. It is also important to point out that calibrating log likelihood ratio with gene length should be taken as the standard procedure in future studies due to the high correlation between them found in our analysis. In addition, in our combined model, length-corrected MCL showed moderate evidence of collinearity with p-distance (Supporting Information Table S3), which suggests that screening for loci with high values of length-corrected MCL may also effectively select loci with other desirable properties.

Chen et al. (2015) proposed a question-specific data filtering strategy to exclude loci that either did not matching predefined hypotheses or did not support a well-established “control node”, which was chosen *a-priori*. They suggested that their methods might work though excluding loci highly subject to random errors or systematic error (Chen et al., 2015). However, when gene tree error is high solely due to loci being short and contain few informative sites, an individual locus may not be expected to strongly support monophyly of a given “control node”, despite potentially still possessing some phylogenetic signal when used in a combined analysis with more loci. Thus, a single node-based topology test may be too conservative for these types of data. Similarly, Arcila et al. (2017) developed a procedure they termed gene genealogy interrogation (GGI), that instead of evaluating each individual gene’s support for a single node, their procedure considers each gene’s support for small subset of predefined alternative topologies in comparison to a single predefined null topological hypothesis. Their study on interfamilial relationships of ostariophysiian fishes showed that, while gene tree error was high in general, only a small subset of genes showed strong support for alternative, unconventional hypotheses, and thus excluding these ‘errant’ genes (putatively paralogous loci) may be an effective filtering strategy (Arcila et al., 2017). Filtering loci based on clocklikeness may also work in a similar way in excluding problematic loci, such as paralogous comparisons that went undetected by other filters, as paralogues will tend to have longer branches and get filtered out by a clocklikeness filter. Unlike the question-specific strategy of Chen et al. (2015) and Arcila et al. (2017), which consider only one node at a time or rely on a small set of predefined phylogenetic hypotheses, filtering based on clocklikeness accounts for tree-wise anomalies and does not require any *a priori* assumptions in phylogenetic relationships among the taxa of interest.

5. Conclusions

New methods in target enrichment gene-capture and advances in next-generation sequencing are generating massive amounts of sequence data with the potential to shed light on some of the most challenging evolutionary questions across the Tree of Life. However, not all data types are informative for most phylogenetic questions and some suites of loci may be inherently biased. There is therefore a need to develop practical ways of selecting phylogenetically useful loci from those that could introduce systematic errors into analyses. Here we used a novel genome-wide phylogenomic dataset of Gobioidei to produce a well-supported phylogeny, which we then utilized to evaluate two potential predictors of phylogenetic performance. We found that molecular clocklikeness is a better indicator of the phylogenetic usefulness of molecular markers in our goby study, and substitution rate was also correlated with phylogenetic accuracy but with low explanatory power. Combining multiple genes with poor characteristics can lead to positively misleading results (strongly-supported, incorrect topologies), regardless of whether these loci are analyzed using coalescent-based species-tree analyses or via a concatenated super-matrix approach. However, our conclusion was based only on the goby dataset, more empirical studies should be performed and studies looking at the utility of different analytical approaches for both species-tree and concatenation analyses (e.g. statistical binning of genes, different partitioning and modeling schemes; Mirarab et al., 2014; Bayzid and Warnow, 2013) will provide more insight into how best to handle phylogenomic data after carefully selecting loci appropriate for a specific phylogenetic question.

Acknowledgements

This work was supported by the Innovation Program of Shanghai Municipal Education Commission; the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning to C. Li. T. Kuang was also supported by Shanghai Outstanding Graduate Scholarship for Interdisciplinary Training. The authors would like to thank Frank Pezold for establishing and generously supporting the collaboration between TAMUCC and SHOU, and for helpful comments on gobioid relationships. The authors also thank Shanghai Oceanus Supercomputing Center (SOSC) for providing computational resource. We have uploaded the fastq files of raw reads (with adapter sequences trimmed) to GenBank (SRR5184353). The assembled sequences can be found in Supporting Information (Supplementary_data_file_3.txt).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ympev.2018.07.018.

References

Agorreta, A., Rüber, L., 2012. A standardized reanalysis of molecular phylogenetic hypotheses of Gobioidei. Syst. Biodivers. 10, 375–390.

Agorreta, A., San Mauro, D., Schliewen, U., Van Tassell, J.L., Kovacic, M., Zardoya, R., Ruber, L., 2013. Molecular phylogenetics of Gobioidei and phylogenetic placement of European gobies. Mol. Phylogenet. Evol. 69, 619–633.

Akihito, Iwata A, Kobayashi, T., Ikeo, K., Imanishi, T., Ono, H., Umehara, Y., Hamamatsu, C., Sugiyama, K., Ikeda, Y., et al., 2000. Evolutionary aspects of gobioid fishes based upon a phylogenetic analysis of mitochondrial cytochrome B genes. Gene 259, 5–15.

Arcila, D., Orti, G., Vari, R., Armbruster, J.W., Stiassny, M.L.J., Ko, K.D., Sabaj, M.H., Lundberg, J., Revell, L.J., Betancur-R, R., 2017. Genome-wide interrogation advances resolution of realcitrant groups in the tree of life. Nat. Ecol. Evol. 1. https://doi.org/10.1038/s41559-016-0020.

Bayzid, M.S., Warnow, T., 2013. Naive binning improves phylogenomic analyses. Bioinformatics 29 (18), 2277–2284.

Betancur, R.R., Broughton, R.E., Wiley, E.O., Carpenter, K., Lopez, J.A., Li, C., Holcroft, N.I., Arcila, D., Sanciangco, M., Cureton Ii, J.C., et al., 2013. The tree of life and a new classification of bony fishes. PLoS Curr. 5.

Birdsong, R.S., Murdy, E.O., Pezold, F.L., 1988. A study of the vertebral column and median fin osteology in gobioid fishes with comments on gobioid relationships. Bull. Mar. Sci. 42, 174–214.

Bossert, S., Murray, E.A., Blaimer, B.B., Danforth, B.N., 2017. The impact of GC bias on phylogenetic accuracy using targeted enrichment phylogenomic data. Mol. Phylogenet. Evol. 111, 149–157.

Bragg, J.G., Potter, S., Bi, K., Moritz, C., 2016. Exon capture phylogenomics: efficacy across scales of divergence. Mol. Ecol. Resour. 16, 1059–1068.

Burleigh, J.G., Bansal, M.S., Eulenstein, O., Hartmann, S., Wehe, A., Vision, T.J., 2011. Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees. Syst. Biol. 60, 117–125.

Camargo, A., Avila, L.J., Morando, M., Sites Jr., J.W., 2012. Accuracy and precision of species trees: effects of locus, individual, and base pair sampling on inference of species trees in lizards of the Liolaemus darwinii group (Squamata, Liolaemidae). Syst. Biol. 61, 272–288.

Capella-Gutierrez, S., Kauff, F., Gabaldon, T., 2014. A phylogenomics approach for selecting robust sets of phylogenetic markers. Nucleic Acids Res. 42, e54.

Chakrabarty, P., Davis, M.P., Sparks, J.S., 2012. The first record of a trans-oceanic sister-group relationship between obligate vertebrate troglobites. PLoS One 7, e44083.

Chen, M.Y., Liang, D., Zhang, P., 2015. Selecting question-specific genes to reduce incongruence in phylogenomics: a case study of jawed vertebrate backbone phylogeny. Syst. Biol. 64, 1104–1120.

Dell'Ampio, E., Meusemann, K., Szucsich, N.U., Peters, R.S., Meyer, B., Borner, J., Petersen, M., Aberer, A.J., Stamatakis, A., Walzl, M.G., et al., 2014. Decisive data sets in phylogenomics: lessons from studies on the phylogenetic relationships of primarily

wingless insects. Mol. Biol. Evol. 31, 239–249.
Dornburg, A., Townsend, J.P., Friedman, M., Near, T.J., 2014. Phylogenetic informativeness reconciles ray-finnsed fish molecular divergence times. BMC Evol. Biol. 14, 169.
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., 2017. New insights on the sister lineage of percomorph fishes with an anchored hybrid enrichment dataset. Mole Phylogenet. Evol. 110, 27–38.
Doyle, V.P., Young, R.E., Naylor, G.J., Brown, J.M., 2015. Can we identify genes with increased phylogenetic reliability? Syst. Biol. 64, 824–837.
Drovetski, S.V., 2002. Molecular phylogeny of grouse: individual and combined performance of W-linked, autosomal, and mitochondrial loci. Syst. Biol. 51, 930–945.
Duchêne, D.A., Duchêne, S., Ho, S.Y.W., 2017. New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. Mol. Biol. Evol. 34 (6), 1529–1534.
Dunn, C.W., Hejnol, A., Matus, D.Q., Pang, K., Browne, W.E., Smith, S.A., Seaver, E., Rouse, G.W., Obst, M., Edgecombe, G.D., et al., 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452, 745–749.
Edwards, S.V., 2016. Phylogenomic subsampling: a brief review. Zool. Scr. 45, 63–74.
Eytan, R.I., Evans, B.R., Dornburg, A., Lemmon, A.R., Lemmon, E.M., Wainwright, P.C., Near, T.J., 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. BMC Evol. Biol. 15, 113.
Faircloth, B.C., McCormack, J.E., Crawford, N.G., Harvey, M.G., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol. 61, 717–726.
Felsenstein, J., 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Seattle, Distributed by the author.
Fiala, K.L., Sokal, R.R., 1985. Factors determining the accuracy of cladogram estimation: evaluation using computer simulation. Evolution 39, 609–622.
Fong, J.J., Fujita, M.K., 2011. Evaluating phylogenetic informativeness and data-type usage for new protein-coding genes across Vertebrata. Mol. Phylogenet. Evol. 61, 300–307.
Frogoso-Martínez, I., Salazar, G.A., Martínez-Gordillo, M., Magallón, S., Sánchez-Reyes, L., Lemmon, E.M., Lemmon, A.R., Sazatornil, F., Granados Mendoza, C., 2017. A pilot study applying the plant Achored Hybrid Enrichment method to New World sages (*Salvia* subgenus *Calosphace*; Lamiaceae). Mol. Phylogenet. Evol. https://doi.org/10.1016/j.ympev.2017.02.006.
Gee, H., 2003. Evolution: ending incongruence. Nature 425, 782.
Hedges, D.J., Guettouche, T., Yang, S., Bademci, G., Diaz, A., Andersen, A., Hulme, W.F., Linker, S., Mehta, A., Edwards, Y.J., et al., 2011. Comparison of three targeted enrichment strategies on the SOLiD sequencing platform. PLoS One 6, e18595.
Hedtke, S.M., Morgan, M.J., Cannatella, D.C., Hillis, D.M., 2013. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. PLoS One 8, e67908.
Hugall, A.F., O'Hara, T.D., Hunjan, S., Nilsen, R., Moussalli, A., 2016. An exon-capture system for the entire class Ophiuroidea. Mol. Biol. Evol. 33, 281–294.
Ilves, K.L., López-Fernández, 2014. A targeted next-generation sequencing toolkit for exon-based cichlid phylogenomics. Mol. Ecol. Resour. 14, 802–811.
Jarvis, E.D., Mirarab, S., Aberer, A.J., Li, B., Houde, P., Li, C., Ho, S.Y., Faircloth, B.C., Nabholz, B., Howard, J.T., et al., 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346, 1320–1331.
Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H., 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22, 225–231.
Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., et al., 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55–61.
Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459–468.
Kumar, S., Filipski, A.J., 2001. Molecular Clock: Testing. eLS.
Lake, J.A., 1991. The order of sequence alignment can bias the selection of tree topology. Mol. Biol. Evol. 8, 378–385.
Lang, J.M., Darling, A.E., Eisen, J.A., 2013. Phylogeny of bacterial and archaeal genomes using conserved genes: supertrees and supermatrices. PLoS One 8, e62510.
Lemmon, A.R., Emme, S.A., Lemmon, E.M., 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol. 61, 727–744.
Li, C., Hofreiter, M., Straube, N., Corrigan, S., Naylor, G.J.P., 2013. Capturing protein-coding genes across highly divergent species. Biotechniques 54, 321–326.
Li, C., Lu, G., Orti, G., 2008. Optimal data partitioning and a test case for ray-finned fishes (actinopterygii) based on ten nuclear loci. Syst. Biol. 57, 519–539.
Li, C., Ortí, G., Zhang, G., Lu, G., 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol. Biol. 7, 44.
Li, C., Ortí, G., 2007. Molecular phylogeny of Clupeiformes (Actinopterygii) inferred from nuclear and mitochondrial DNA sequences. Mol. Phylogenet. Evol. 44, 386–398.
Li, C., Riethoven, J.J., Naylor, G.J.P., 2012. EvolMarkers: a database for mining exon and intron markers for evolution, ecology and conservation studies. Mol. Ecol. Resour. 12, 967–971.
Liu, K., Linder, C.R., Warnow, T., 2010. Multiple sequence alignment: a major challenge to large-scale phylogenetics. PLoS Curr. 2, RRN1198.
Longo, G., Bernardi, G., 2015. The evolutionary history of the embiotocid surfperch radiation based on genome-wide RAD sequence data. Mol. Phylogenet. Evol. 88, 55–63.
Lopez-Giraldez, F., Moeller, A.H., Townsend, J.P., 2013. Evaluating phylogenetic informativeness as a predictor of phylogenetic signal for metazoan, fungal, and mammalian phylogenomic data sets. Biomed Res Int. 2013, 621604.
McCormack, J.E., Faircloth, B.C., Crawford, N.G., Gowaty, P.A., Brumfield, R.T., Glenn, T.C., 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Res. 22, 746–754.
Mirarab, S., Bayzid, M.S., Boussau, B., Warnow, T., 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346, 1250463.
Mirarab, S., Warnow, T., 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds of taxa and thousands of genes. Bioinformatics 31 (12), i44–i52.
Muse, S.V., Weir, B.S., 1992. Testing for equality of evolutionary rates. Genetics 132, 269–276.
Nosenko, T., Schreiber, F., Adamska, M., Adamski, M., Eitel, M., Hammel, J., Maldonado, M., Muller, W.E., Nickel, M., Schierwater, B., et al., 2013. Deep metazoan phylogeny: when different genes tell different stories. Mol. Phylogenet. Evol. 67, 223–233.
Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T., Manuel, M., Worheide, G., Baurain, D., 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.
Philippe, H., Derelle, R., Lopez, P., Pick, K., Borchiellini, C., Boury-Esnault, N., Vacelet, J., Renard, E., Houliston, E., Queinnec, E., et al., 2009. Phylogenomics revives traditional views on deep animal relationships. Curr. Biol. 19, 706–712.
Phillips, M.J., Delsuc, F., Penny, D., 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21, 1455–1458.
Phillips, M.J., Penny, D., 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogenet. Evol. 28, 171–185.
Prum, R.O., Berv, S.J., Dornburg, A., Filed, D.J., Townsend, J.P., Lemmon, E.M., Lemmon, A.R., 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526, 569–573.
Qiao, B., Goldberg, T.L., Olsen, G.J., Weigel, R.M., 2006. A computer simulation analysis of the accuracy of partial genome sequencing and restriction fragment analysis in the reconstruction of phylogenetic relationships. Infect Genet. Evol. 6, 323–330.
Qu, X.-J., Jin, J.-J., Chaw, S.-M., Li, D.-Z., Yi, T.-S., 2017. Multiple measures could alleviate long-branch attraction in phylogenomic reconstruction of the Cupressoideae (Cupressaceae). Sci. Rep. 7, 41005. https://doi.org/10.1038/srep41005.
Ramsey, J.B., 1969. Tests for specification errors in classica linear least squares regression analysis. J. Roy. Stat. Soc. B 31, 250–371.
Robinson, D., Foulds, L., 1981. Comparison of phylogenetic trees. Math. Biosci. 53.
Rodriguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389–399.
Rokas, A., Carroll, S.B., 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22, 1337–1344.
Roure, B., Baurain, D., Philippe, H., 2013. Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Mol. Biol. Evol. 30, 197–214.
Rüber, L., Agorreta, A., 2011. Molecular systematics of gobioid fishes. In: Patzner, R.A., Van Tassell, J.L., Kovacic, M., Kapoor, B.G. (Eds.), The Biology of Gobies. Science Publishers, Enfield, (NH), pp. 23–50.
Salichos, L., Rokas, A., 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497, 327–331.
Schierwater, B., Eitel, M., Jakob, W., Osigus, H.J., Hadrys, H., Dellaporta, S.L., Kolokotronis, S.O., Desalle, R., 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis. PLoS Biol. 7, e20.
Sharma, P.P., Kaluziak, S.T., Pérez-Porro, A.R., González, V.L., Hormiga, G., Wheeler, W.C., Giribet, G., 2014. Phylogenomic interrogation of Arachnida reveals systematic conflicts in phylogenetic signal. Mol. Biol. Evol. 31, 2963–2984. https://doi.org/10.1093/molbev/msu235.
Sayyari, E., Mirarab, S., 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33 (7), 1654–1668.
Stamatakis, A., Hoover, P., Rougemont, J., 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst. Biol. 57, 758–771.
Stamatakis, A., Ludwig, T., Meier, H., 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21, 456–463.
Stamatakis, A., 2006. Phylogenetic models of rate heterogeneity: a high performance computing perspective. In: Proc. of IPDPS2006, Rhodos, Greece.
Streicher, J.W., Schulte, J.A., Wiens, J.J., 2016. How should genes and taxa be sampled for phylogenomic analyses with missing data? An empirical study in iguanian lizards. Syst. Biol. 65, 128–145.
Swofford, D.L., 2000. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
Talavera, G., Castresana, J., 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56, 564–577.
Tewhey, R., Nakano, M., Wang, X., Pabon-Pena, C., Novak, B., Giuffre, A., Lin, E., Happe, S., Roberts, D.N., LeProust, E.M., et al., 2009. Enrichment of sequencing targets from the human genome by solution hybridization. Genome Biol. 10, R116.
Thacker, C.E., Hardman, M.A., 2005. Molecular phylogeny of basal gobioid fishes: Rhyacichthyidae, Odontobutidae, Xenisthmidae, Eleotridae (Teleostei: Perciformes: Gobioidei). Mol. Phylogenet. Evol. 37, 858–871.
Thacker, C.E., Roje, D.M., 2011. Phylogeny of Gobiidae and identification of gobiid lineages. Syst. Biodivers. 9, 329–347.
Thacker, C.E., Satoh, T.P., Katayama, E., Harrington, R.C., Eytan, R.I., Near, J., 2015. Molecular phylogeny of Percomorpha resolves Trichonotus as the sister lineage to Gobioidei (Teleostei: Gobiiformes) and confirms the polyphyly of Trachinoidei. Mol. Phylogenet. Evol. 93, 172–179.
Thacker, C.E., 2003. Molecular phylogeny of the gobioid fishes (Teleostei: Perciformes: Gobioidei). Mol. Phylogenet. Evol. 26, 354–368.
Thacker, C.E., 2009. Phylogeny of Gobioidei and placement within Acanthomorpha, with a new classification and investigation of diversification and character evolution. Copeia 2009, 93–104.
Thacker, C.E., 2013. Phylogenetic placement of the European sand gobies in

Gobionellidae and characterization of gobionellid lineages (Gobiiformes: Gobioidei). Zootaxa 3619, 369–382.

Thacker, C.E., 2014. Species and shape diversification are inversely correlated among gobies and cardinalfishes (Teleostei: Gobiiformes). Organismal Divers. Evol. 14, 419–436.

Tornabene, L., Chen, Y., Pezold, F., 2013. Gobies are deeply divided: phylogenetic evidence from nuclear DNA (Teleostei: Gobioidei: Gobiidae). Syst. Biodivers. 2013, 1–17.

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

Van Tassell, J., Tornabene, L., Taylor, M., 2011. A history of gobioid morphological systematics. In: Patzner, R., Van Tassell, J., Kovacic, M., Kapoor, B. (Eds.), The Biology of Gobies. Science Publishers Inc., Enfield, NH, pp. 3–22.

Winterbottom, R., 1993. Search for the gobioid sister group (Actinopterygii: Percomorpha). Bull. Mar. Sci. 52, 395–414.